

*IBM SPSS Modeler 16 - Nodi di origine,
elaborazione e output*

IBM

Nota

Prima di utilizzare queste informazioni e il relativo prodotto, leggere le informazioni in "Note" a pagina 359.

Informazioni sul prodotto

Questa edizione si applica alla versione 16, release 0, modifica 0 di IBM(r) SPSS(r) Modeler e a tutte le release e modifiche successive, se non diversamente indicato nelle nuove edizioni.

Indice

Prefazione	vii
-----------------------------	------------

Capitolo 1. Informazioni su IBM SPSS Modeler **1**

Prodotti IBM SPSS Modeler	1
IBM SPSS Modeler	1
IBM SPSS Modeler Server	1
IBM SPSS Modeler Administration Console	2
IBM SPSS Modeler Batch	2
IBM SPSS Modeler Solution Publisher	2
Adattatori IBM SPSS Modeler Server per IBM SPSS Collaboration and Deployment Services	2
Edizioni di IBM SPSS Modeler	2
Documentazione di IBM SPSS Modeler	3
Documentazione di SPSS Modeler Professional	3
Documentazione di SPSS Modeler Premium	4
Esempi di applicazioni	5
Cartella Demos	5

Capitolo 2. Nodi origine **7**

Osservazioni generali	7
Nodo Enterprise View	8
Impostazione delle opzioni relative al nodo Enterprise View	9
Connessioni Enterprise View	10
Scelta della Definizione dei provider di dati	11
Scelta della tabella	11
Nodo Vista dati	11
Impostazione delle opzioni relative al nodo Vista dati	12
Nodo origine del database	13
Impostazione delle opzioni relative al nodo Database	14
Aggiunta di una connessione a un database	15
Specificazione di valori preimpostati per una connessione a un database	15
Selezione di una tabella di database	18
Invio di query al database	18
Nodo Testo variabile	20
Impostazione delle opzioni relative al nodo Testo variabile	20
Nodo Testo fisso	22
Impostazione delle opzioni relative al nodo Testo fisso	22
Impostazione dell'archiviazione e della formattazione dei campi	24
Nodo Data Collection	25
Opzioni scheda File del nodo Importazione di Data Collection	26
Proprietà dei metadati dell'importazione di IBM SPSS Data Collection	28
Stringa di connessione al database	29
Proprietà avanzate	29
Importazione di insiemi di risposte multiple	29

Note sull'importazione di colonne in IBM SPSS Data Collection	29
Analytic ServerOrigine	30
Selezione origine dati di Analytic Server.	30
Credenziali di Analytic Server	31
Nodi supportati	31
Nodo origine IBM Cognos BI	34
Icone oggetti Cognos	35
Importazione di dati Cognos	36
Importazione di report Cognos	36
Connessioni Cognos	37
Selezione del percorso Cognos	37
Specificazione dei parametri per i dati o i report.	38
Nodo origine IBM Cognos TM1	38
Importazione di dati Cognos TM1.	38
Nodo origine SAS	39
Impostazione delle opzioni relative al nodo origine SAS	39
Nodo origine Excel	40
Nodo origine XML	41
Selezione da più elementi root	42
Rimozione degli spazi indesiderati dai dati di origine XML	42
Nodo Genera simulazione	42
Impostazione delle opzioni relative al nodo Genera simulazione	44
Clona campo	49
Dettagli adattamento	50
Specificazione parametri	51
Distribuzioni	53
Nodo input utente	55
Impostazioni relative al nodo Input utente	56
Schede comuni a tutti i nodi origine	59
Impostazione dei livelli di misurazione nel nodo origine	59
Filtro di campi dal nodo origine	60

Capitolo 3. Nodi Operazioni su record **61**

Panoramica sui nodi Operazioni su record	61
Nodo Seleziona	62
Nodo Campione	63
Opzioni del nodo Campione	64
Impostazioni di raggruppamento e stratificazione	66
Dimensioni dei campioni per gli strati	67
Nodo bilanciamento	68
Impostazione delle opzioni per il nodo bilanciamento	68
Nodo Aggregazione	69
Impostazione delle opzioni per il nodo Aggregazione.	69
Nodo Aggregazione RFM.	72
Impostazione delle opzioni per il nodo Aggregazione RFM.	72
Nodo Ordina	73
Impostazioni di ottimizzazione dell'ordinamento	73

Nodo Unione	74	Impostazione delle opzioni relative al nodo Crea variabile conteggio	131
Tipi di unioni	74	Impostazione delle opzioni relative al nodo Ricava condizionale	131
Definizione di chiavi e metodo di unione	76	Ricodifica di valori con il nodo Ricava	132
Selezione di dati per unioni parziali	77	Nodo Riempimento	132
Specifica delle condizioni per un'unione	77	Conversione della classe di archiviazione mediante il nodo Riempimento	133
Filtro di campi dal nodo Unione	77	Nodo Anonimizza	133
Impostazione di tag e ordine dei dati di input	78	Impostazione delle opzioni relative al nodo Anonimizza	134
Impostazioni di ottimizzazione dell'unione	79	Anonimizzazione dei valori dei campi	135
Nodo Accodamento	80	Nodo Ricodifica	136
Impostazione delle opzioni relative al nodo Accodamento	80	Impostazione delle opzioni per il nodo Ricodifica	136
Nodo Elimina duplicati	80	Ricodifica di più campi	137
Impostazioni di ottimizzazione del nodo Elimina duplicati	82	Classe di archiviazione e livello di misurazione per i campi ricodificati	137
Impostazioni di Composito distinto	83	Nodo Discretizza	138
Nodo Streaming TS	85	Impostazione delle opzioni relative al nodo Discretizza	139
Opzioni dei campi di Streaming TS	86	Bin a larghezza fissa	139
Opzioni del modello Streaming TS	86	Blocchi (conteggio o somma uguale).	140
Criteri di Expert Modeler per le serie temporali	87	Classifica casi	141
Criteri di livellamento esponenziale per le serie temporali	88	Deviazione standard/Media	142
Criteri ARIMA per le serie temporali	89	Categorizzazione ottimale	142
Funzioni di trasferimento	90	Anteprima dei bin generati	143
Gestione dei valori anomali	91	Nodo Analisi RFM	144
Opzioni di distribuzione di Streaming TS	91	Impostazioni del nodo Analisi RFM	144
Nodo STB (Space-Time-Box)	92	Discretizzazione con il nodo Analisi RFM	145
Definizione della densità STB (Space-Time-Box)	93	Nodo Partizione	146
Capitolo 4. Nodi Operazioni su campi	95	Opzioni del nodo Partizione	146
Panoramica sui nodi Operazioni sui campi	95	Nodo Crea flag	147
Preparazione automatica dati	97	Impostazione delle opzioni relative al nodo Crea flag	147
Scheda Campi	98	Nodo Riorganizza	148
Scheda Impostazioni	99	Impostazione delle opzioni relative al nodo Riorganizza	149
Scheda Analisi	104	Nodo Trasponi	150
Generazione di un nodo Ricava	110	Impostazione delle opzioni del nodo Trasponi	150
Nodo Tipo	111	Nodo Intervalli di tempo	151
Livelli di misurazione	112	Indicazione degli intervalli di tempo	152
Conversione di dati continui	114	Opzioni di creazione degli intervalli di tempo	153
Che cos'è la creazione istanza?	114	Periodo di stima	154
Valori dei dati	115	Previsioni	154
Definizione dei valori mancanti	118	Intervalli supportati	155
Controllo dei valori del tipo	119	Nodo Cronologia	162
Impostazione del ruolo del campo	119	Impostazione delle opzioni relative al nodo Cronologia	162
Copia degli attributi del tipo	120	Nodo Ordina campi	163
Scheda per l'impostazione dei formati dei campi	120	Impostazione delle opzioni Ordina campi	163
Applicazione di filtri o ridenominazione di campi	122	Capitolo 5. Nodi Grafico	165
Impostazione delle opzioni di filtro	122	Caratteristiche comuni dei nodi Grafico	165
Nodo dell'insieme	125	Effetti estetici, sovrapposizioni, riquadri e animazioni	166
Impostazioni del nodo dell'insieme	126	Utilizzo della scheda Output	167
Nodo Ricava	127	Utilizzo della scheda Annotazioni.	168
Impostazione delle opzioni di base relative al nodo Crea variabile	128	Grafici 3-D	168
Derivazione di campi multipli	128	Nodo Lavagna grafica	170
Impostazione delle opzioni relative al nodo Crea variabile formula	129	Scheda Di base del nodo Lavagna grafica	170
Impostazione delle opzioni relative al nodo Ricava indicatore	130		
Impostazioni delle opzioni del nodo nuovo campo nominale	130		
Impostazione delle opzioni relative al nodo Ricava stato	131		

Lavagna grafica Scheda Dettagliato	174
Tipi di visualizzazione della lavagna grafica incorporati disponibili	176
Creazione di visualizzazioni delle mappe	183
Lavagna grafica Esempi	184
Scheda Aspetto del nodo Lavagna grafica	194
Impostazione della posizione di modelli, fogli di stile e mappe	196
Gestione dei modelli, dei fogli di stile e dei file delle mappe	196
Conversione e distribuzione degli shapefile delle mappe.	197
Concetti principali delle mappe	198
Utilizzo del programma di utilità di conversione mappe.	199
Distribuzione dei file mappa	204
Nodo Plot	205
Scheda Nodo Plot	207
Scheda Opzioni del nodo Plot	208
Scheda Aspetto del nodo Plot	210
Utilizzo del grafico	210
Nodo distribuzione	211
Scheda Plot del nodo distribuzione	211
Scheda Aspetto del nodo distribuzione	212
Utilizzo di un nodo distribuzione	212
Nodo Istogramma	215
Scheda Plot del nodo Istogramma	215
Scheda Opzioni del nodo Istogramma	215
Scheda Aspetto del nodo Istogramma	216
Utilizzo degli istogrammi	216
Nodo Raccolta	217
Scheda Plot del nodo Raccolta	217
Scheda Opzioni del nodo Raccolta	218
Scheda Aspetto del nodo Raccolta	218
Utilizzo del grafico raccolta.	219
Nodo Multiplot	220
Scheda Grafico di un nodo Multiplot	220
Scheda Aspetto del nodo Multiplot	222
Utilizzo del grafico multiplot	222
Nodo Web	223
Scheda Plot del nodo Web	224
Scheda Opzioni del nodo Web.	225
Scheda Aspetto del nodo Web	226
Utilizzo di un grafico Web	227
Nodo del grafico temporale	231
Scheda Plot del nodo Plot tempo	231
Scheda Aspetto del nodo Plot tempo	232
Utilizzo del grafico tempo	233
Nodo Valutazione	233
Scheda Plot del nodo Valutazione	238
Scheda Opzioni del nodo Valutazione	240
Scheda Aspetto del nodo Valutazione	240
Lettura dei risultati di un grafico di valutazione relativo a un modello.	241
Utilizzo di un grafico di valutazione	242
Esplorazione dei grafici	243
Utilizzo delle sezioni	244
Utilizzo delle aree	248
Utilizzo degli elementi contrassegnati	250
Generazioni di nodi da grafici	251
Modifica delle visualizzazioni	254

Regole generali per la modifica delle visualizzazioni	255
Modifica e formattazione del testo	256
Modifica dei colori, dei motivi, dei tratteggi e della trasparenza	256
Rotazione e modifica della forma e delle proporzioni degli elementi puntiformi	257
Modifica delle dimensioni degli elementi grafici	257
Impostazione dei margini e del padding	258
Formattazione dei numeri	258
Modifica delle impostazioni degli assi e delle scale	259
Modifica delle categorie	261
Modifica dei riquadri di orientamento	262
Trasformazione del sistema di coordinate	262
Modifica di statistiche ed elementi grafici	263
Modifica della posizione della legenda	266
Copia di una visualizzazione e dei dati delle visualizzazioni	266
Tasti di scelta rapida	266
Aggiunta di titoli e annotazioni	266
Utilizzo dei fogli di stile dei grafici	267
Stampa, salvataggio, copia ed esportazione di grafici	268

Capitolo 6. Nodi Output 271

Panoramica sui nodi Output	271
Gestione dell'output	272
Visualizzazione dell'output	272
Pubblica sul Web	273
Visualizzazione dell'output in un browser HTML.	274
Esportazione dell'output.	274
Selezione di celle e di colonne.	275
Nodo Tabella	275
Scheda Impostazioni del nodo Tabella	275
Scheda Formato del nodo Tabella.	275
Scheda Output del nodo Output	276
Browser del nodo Tabella	277
Nodo Matrice	278
Scheda Impostazioni del nodo Matrice	278
Scheda Aspetto del nodo Matrice.	279
Browser di output del nodo Matrice.	279
nodo Analisi.	280
Scheda Analisi del nodo Analisi	280
Browser di output del nodo Analisi	282
Nodo Esplora	284
Scheda Impostazioni del nodo Esplora	284
Nodo Esplora, scheda Qualità	285
Browser di output del nodo Esplora.	285
Nodo Trasformazioni.	291
Scheda Opzioni del nodo Trasformazioni	291
Scheda Output del nodo Trasformazioni	292
Visualizzatore output del nodo Trasformazioni	292
Nodo Statistiche	294
Scheda Impostazioni del nodo Statistiche	294
Browser di output del nodo Statistiche	295
Nodo Medie.	296
Confronto di medie per gruppi indipendenti	296
Confronto delle medie tra campi accoppiati	297
Opzioni del nodo Medie.	297

Browser di output del nodo Medie	297
Nodo Report	299
Scheda Modello del nodo Report	299
Browser di output del nodo Report	300
Nodo Calcola globali	300
Scheda Impostazioni del nodo Calcola globali	301
Nodo Adattamento simulazione	301
Adattamento delle distribuzioni	302
Scheda Impostazioni del nodo Adattamento simulazione	303
Nodo Valutazione simulazione	304
Scheda Impostazioni del nodo Valutazione simulazione	304
Output del nodo Valutazione simulazione . . .	306
IBM SPSS Statistics Applicazioni di supporto . .	312

Capitolo 7. Nodi di esportazione 315

Panoramica sui nodi di esportazione	315
Nodo di esportazione del database	316
Scheda di esportazione del nodo Database . . .	316
Opzioni di unione esportazione database . . .	317
Esportazione Database - Opzioni Schema . . .	318
Esportazione database - Opzioni Indice . . .	321
Esportazione database - Opzioni avanzate . . .	323
Programmi di caricamento di massa	324
Nodo di esportazione File flat	331
Scheda di esportazione del nodo File flat . . .	331
Nodo Esporta IBM SPSS Data Collection	331
Esportazione di Analytic Server	332
Nodo di esportazione IBM Cognos BI	333
Connessione Cognos	333
Connessione ODBC	334
Nodo di esportazione IBM Cognos TM1	335
Connessione a un cubo Cognos TM1 per esportare i dati	335
Mapping dei dati Cognos TM1 per l'esportazione	336
Nodo Esporta SAS	336
Scheda Esporta del nodo Esporta SAS	336
Nodo Esportazione da Excel	337
Scheda di esportazione del nodo Excel	337
Nodo Esporta XML	337
Scrittura dei dati XML	338
Opzioni dei record di mappatura XML	339
Opzioni dei campi di mappatura XML	339

Anteprima mappatura XML	339
-----------------------------------	-----

Capitolo 8. Nodi IBM SPSS Statistics 341

Nodi IBM SPSS Statistics - Panoramica	341
Nodo File Statistics	342
Nodo Trasformazioni Statistics	343
Scheda Sintassi del nodo Trasformazioni Statistics	343
Sintassi consentita	344
Nodo Modello Statistics	346
Scheda Modello del nodo Modello Statistics . .	346
Riepilogo del nugget del modello del nodo Modello Statistics	346
Nodo Output Statistics	346
Scheda Sintassi del nodo Output Statistics . . .	347
Scheda Output del nodo Output Statistics . . .	348
Nodo Esporta Statistics	349
Scheda Esporta del nodo Esporta Statistics . .	349
Ridenominazione o filtro dei campi per IBM SPSS Statistics	350

Capitolo 9. Supernodi 351

Panoramica sui Supernodi	351
Tipi di Supernodi	351
Supernodi origine	351
Supernodi di elaborazione	351
Supernodo finale	352
Creazione di Supernodi	352
Nidificazione di Supernodi	353
Blocco di Supernodi	353
Blocco e sblocco di Supernodi	353
Modifica di un Supernodo bloccato	354
Modifica Supernodo	354
Modifica dei tipi di Supernodo	354
Annotazione e ridenominazione dei Supernodi	355
Parametri dei Supernodi	355
Supernodi e cache	357
Supernodi e script	357
Salvataggio e caricamento dei Supernodi	358

Note 359

Marchi	360
------------------	-----

Indice analitico. 365

Prefazione

IBM® SPSS Modeler è l'efficace workbench di data mining aziendale di IBM Corp.. SPSS Modeler consente alle organizzazioni di migliorare le relazioni con i clienti e con il pubblico grazie a un'analisi approfondita dei dati. Le organizzazioni potranno utilizzare le informazioni ottenute tramite SPSS Modeler per mantenere i clienti di valore, cogliere opportunità di vendite incrociate, attrarre nuovi clienti, individuare frodi, diminuire i rischi e migliorare l'offerta di servizi a livello statale.

L'interfaccia visuale di SPSS Modeler invita gli utenti ad applicare le proprie competenze di business specifiche, consentendo di ottenere modelli previsionali più potenti e di ridurre i tempi necessari per lo sviluppo di soluzioni. SPSS Modeler offre molte tecniche di modellazione, come previsione, classificazione, segmentazione e algoritmi di rilevamento delle associazioni. IBM SPSS Modeler Solution Publisher consente quindi di distribuire a livello aziendale i modelli creati in modo che vengano utilizzati dai responsabili dei processi decisionali oppure inseriti in un database.

Informazioni su IBM Business Analytics

Il software IBM Business Analytics fornisce informazioni complete, coerenti e accurate a cui i responsabili delle decisioni possono affidarsi per ottimizzare le prestazioni di business. Un ampio portafoglio di applicazioni di business intelligence, analisi predittiva, gestione delle prestazioni e delle strategie finanziarie e analisi offre una panoramica chiara, istantanea e interattiva delle prestazioni attuali e la possibilità di prevedere i risultati futuri. Utilizzato in combinazione con potenti soluzioni di settore, prassi consolidate e servizi professionali, questo software consente alle aziende di tutte le dimensioni di ottimizzare la produttività, automatizzare le decisioni senza problemi e fornire risultati migliori.

Come parte di questo portafoglio, il software IBM SPSS Predictive Analytics consente alle aziende di prevedere gli eventi futuri e di agire tempestivamente in modo da migliorare i risultati delle attività di business. Le aziende, gli enti governativi e le università di tutto il mondo si affidano alla tecnologia IBM SPSS perché rappresenta un vantaggio concorrenziale in termini di attrazione, retention e aumento dei clienti, riducendo al tempo stesso, le frodi e limitando i rischi. Incorporando il software IBM SPSS nelle operazioni quotidiane, le organizzazioni diventano imprese in grado di effettuare previsioni - in grado di dirigere e rendere automatiche le decisioni per raggiungere gli obiettivi aziendali ed acquisire vantaggi tangibili rispetto alla concorrenza. Per ulteriori informazioni o per contattare un rappresentante, visitare il sito <http://www.ibm.com/spss>.

Supporto tecnico

Il supporto tecnico è a disposizione dei clienti che dispongono di un contratto di manutenzione. I clienti possono contattare il supporto tecnico per richiedere assistenza per l'utilizzo di IBM Corp. o per l'installazione di uno degli ambienti hardware supportati. Per rivolgersi al supporto tecnico, consultare il sito Web IBM Corp. all'indirizzo <http://www.ibm.com/support>. Ricordare che durante la richiesta di assistenza sarà necessario fornire i dati di identificazione personali, i dati relativi alla propria società e il numero del accordo di manutenzione.

Capitolo 1. Informazioni su IBM SPSS Modeler

IBM SPSS Modeler è un insieme di strumenti di data mining che consente di sviluppare rapidamente modelli predittivi con l'ausilio di competenze di business e di eseguirne la distribuzione nelle operazioni di business per migliorare i processi decisionali. Progettato secondo il modello CRISP-DM conforme agli standard di settore, IBM SPSS Modeler supporta l'intero processo di data mining, dai dati a risultati di business migliori.

IBM SPSS Modeler offre numerosi metodi di modellazione ricavati dall'apprendimento automatico, dall'intelligenza artificiale e dalla statistica. I metodi disponibili nella palette Modelli consentono di ricavare nuove informazioni dai dati e di sviluppare modelli predittivi. Ogni metodo ha determinati punti di forza e si presta meglio per particolari tipi di problemi.

SPSS Modeler può essere acquistato come prodotto autonomo oppure utilizzato come client in combinazione con SPSS Modeler Server. È inoltre disponibile una serie di opzioni, come illustrato nelle sezioni seguenti. Per ulteriori informazioni, consultare <http://www.ibm.com/software/analytics/spss/products/modeler/>.

Prodotti IBM SPSS Modeler

La famiglia di prodotti IBM SPSS Modeler e del software associato comprende quanto segue.

- IBM SPSS Modeler
- IBM SPSS Modeler Server
- IBM SPSS Modeler Administration Console
- IBM SPSS Modeler Batch
- IBM SPSS Modeler Solution Publisher
- Adattatori IBM SPSS Modeler Server per IBM SPSS Collaboration and Deployment Services

IBM SPSS Modeler

SPSS Modeler è una versione del prodotto con funzionalità complete che viene installata ed eseguita sul proprio PC. È possibile eseguire SPSS Modeler in modalità locale come prodotto autonomo oppure in modalità distribuita assieme a IBM SPSS Modeler Server per ottenere una migliore performance su insiemi di dati di grandi dimensioni.

Grazie a SPSS Modeler si possono creare, in modo veloce e intuitivo, modelli predittivi accurati senza ricorrere alla programmazione. La sua avanzata interfaccia visiva permette di visualizzare con facilità il processo di data mining. Grazie alle funzionalità di analisi avanzate incorporate nel prodotto, l'utente potrà rilevare la presenza di pattern e tendenze, che altrimenti rimarrebbero occulti, all'interno dei dati. La modellazione dei risultati e la comprensione dei fattori che li influenzano consente di beneficiare di maggiori opportunità di business e, al contempo, di ridurre i rischi.

SPSS Modeler è disponibile in due edizioni: SPSS Modeler Professional e SPSS Modeler Premium. Per ulteriori informazioni, consultare l'argomento "Edizioni di IBM SPSS Modeler" a pagina 2.

IBM SPSS Modeler Server

SPSS Modeler utilizza un'architettura client/server per distribuire le richieste di operazioni che utilizzano molte risorse a potenti componenti software server, con un conseguente miglioramento della performance su insiemi di dati di grandi dimensioni.

SPSS Modeler Server è un prodotto con licenza separata che viene eseguito continuamente in modalità di analisi distribuita su un host server insieme a una o più installazioni IBM SPSS Modeler. Una configurazione di questo tipo consente a SPSS Modeler Server di ottenere prestazioni migliori quando si lavora su insiemi di dati di grandi dimensioni, in quanto le operazioni che richiedono un utilizzo consistente della memoria possono essere eseguite sul server senza scaricare i dati sul computer client. IBM SPSS Modeler Server offre inoltre il supporto delle funzionalità di ottimizzazione SQL e di modellazione nel database, garantendo ulteriori benefici dal punto di vista delle prestazioni e del livello di automazione.

IBM SPSS Modeler Administration Console

Modeler Administration Console è un'applicazione grafica per la gestione di molte delle opzioni di configurazione di SPSS Modeler Server, la cui configurazione può avvenire, inoltre, mediante un file delle opzioni. L'applicazione fornisce un'interfaccia utente di console per monitorare e configurare le installazioni di SPSS Modeler Server ed è disponibile gratuitamente per i clienti esistenti di SPSS Modeler Server. L'applicazione può essere installata solo sui computer Windows; tuttavia, può gestire un server installato su qualsiasi piattaforma supportata.

IBM SPSS Modeler Batch

Nonostante il data mining sia generalmente un processo di tipo interattivo, è possibile eseguire SPSS Modeler da una riga di comando senza il bisogno di ricorrere all'interfaccia utente grafica. Poniamo, ad esempio, che si debbano svolgere varie attività laboriose e ripetitive che non richiedono l'intervento di un utente. SPSS Modeler Batch è una versione speciale del prodotto che supporta l'intera gamma di funzionalità analitiche di SPSS Modeler senza richiedere l'accesso all'interfaccia utente normale. Per utilizzare SPSS Modeler Batch, è necessario disporre di una licenza SPSS Modeler Server.

IBM SPSS Modeler Solution Publisher

SPSS Modeler Solution Publisher è uno strumento che consente di creare una versione a pacchetto di un flusso SPSS Modeler che potrà essere eseguito da un motore di runtime esterno oppure incorporato in una applicazione esterna. Questo permette di pubblicare e sottoporre a deployment stream SPSS Modeler completi in ambienti in cui SPSS Modeler non è installato. SPSS Modeler Solution Publisher è distribuito come parte del servizio IBM SPSS Collaboration and Deployment Services - Scoring, per cui è necessario procurarsi una licenza separata. Insieme alla licenza, si riceve SPSS Modeler Solution Publisher Runtime, che consente di eseguire i flussi pubblicati.

Adattatori IBM SPSS Modeler Server per IBM SPSS Collaboration and Deployment Services

È disponibile una serie di adattatori per IBM SPSS Collaboration and Deployment Services che abilitano l'interazione di SPSS Modeler e SPSS Modeler Server con un repository IBM SPSS Collaboration and Deployment Services. In questo modo, un flusso SPSS Modeler sottoposto a deployment sul repository potrà essere condiviso da più utenti oppure risulterà accessibile dall'applicazione thin client IBM SPSS Modeler Advantage. L'adattatore va installato sul sistema che ospita il repository.

Edizioni di IBM SPSS Modeler

SPSS Modeler è disponibile nelle edizioni seguenti.

SPSS Modeler Professional

SPSS Modeler Professional contiene tutti gli strumenti necessari per utilizzare la maggior parte dei tipi di dati strutturati, quali comportamenti e interazioni registrati in sistemi CRM, dati demografici, dati sulle vendite e sul comportamento d'acquisto.

SPSS Modeler Premium

SPSS Modeler Premium è un prodotto con licenza separata che amplia l'ambito di utilizzo di SPSS Modeler Professional aggiungendo il supporto di dati speciali, quali quelli usati per l'analisi delle entità o dei social network, e di dati di testo non strutturati. SPSS Modeler Premium comprende i seguenti componenti.

IBM SPSS Modeler Entity Analytics aggiunge una dimensione supplementare alle analisi predittive di IBM SPSS Modeler . Se l'analisi predittiva tenta di prevedere il comportamento futuro sulla base di dati precedenti, l'analisi dell'entità si concentra sul miglioramento della coerenza dei dati correnti risolvendo i conflitti tra gli stessi record. Un'identità può essere di un individuo, un'organizzazione, un oggetto o qualsiasi altra entità per cui possa esistere ambiguità. La risoluzione dell'identità può essere essenziale in diversi campi, tra cui la gestione delle relazioni con i clienti, il rilevamento di frodi, il riciclaggio di denaro e la sicurezza nazionale e internazionale.

IBM SPSS Modeler Social Network Analysis trasforma le informazioni sulle relazioni in campi che caratterizzano il comportamento sociale di individui e gruppi. Facendo leva sui dati che descrivono le relazioni esistenti nelle reti sociali, IBM SPSS Modeler Social Network Analysis riesce a individuare i leader in grado di influenzare il comportamento degli altri membri della rete. Consente inoltre di stabilire quali individui della rete sono maggiormente influenzati dagli altri membri. La combinazione di questi risultati ad altre misurazioni permette di delineare profili complessi degli individui su cui basare dei modelli predittivi. I modelli che contengono informazioni sociali generano risultati più accurati rispetto agli altri.

IBM SPSS Modeler Text Analytics utilizza tecnologie linguistiche avanzate e di NLP (Natural Language Processing) per elaborare rapidamente una grande varietà di dati di testo non strutturati, estrarre ed organizzare i concetti chiave e raggruppare tali concetti in categorie. È quindi possibile combinare i concetti e le categorie estratti con dati strutturati esistenti, per esempio dati demografici, e applicarli alla modellazione utilizzando la suite completa degli strumenti di data mining di IBM SPSS Modeler per prendere decisioni migliori e più mirate.

Documentazione di IBM SPSS Modeler

La documentazione nel formato guida in linea è disponibile nel menu Aiuto di SPSS Modeler. Sono incluse la documentazione per SPSS Modeler, SPSS Modeler Server e SPSS Modeler Solution Publisher, nonché la Guida alle applicazioni e altro materiale di supporto.

La documentazione completa in formato PDF dei singoli prodotti, istruzioni di installazione comprese, è disponibile nella cartella *\Documentation* del DVD di ciascun prodotto. I documenti per l'installazione possono anche essere scaricati dal Web, all'indirizzo <http://www-01.ibm.com/support/docview.wss?uid=swg27038316>.

La documentazione in entrambi i formati è inoltre disponibile presso il Centro informazioni SPSS Modeler all'indirizzo <http://publib.boulder.ibm.com/infocenter/spssmodl/v16r0m0/>.

Documentazione di SPSS Modeler Professional

La documentazione completa di SPSS Modeler Professional, escluse le istruzioni di installazione, è la seguente.

- **IBM SPSS Modeler - Guida per l'utente.** Introduzione generale all'utilizzo di SPSS Modeler che illustra come creare flussi di dati, gestire valori mancanti, generare espressioni CLEM, utilizzare progetti e report e assemblare flussi per la distribuzione tramite IBM SPSS Collaboration and Deployment Services, le applicazioni predittive o IBM SPSS Modeler Advantage.
- **IBM SPSS Modeler - Nodi origine, elaborazione e output.** Descrizioni di tutti i nodi utilizzati per leggere, elaborare e generare dati di output in vari formati, ovvero di nodi ad eccezione dei nodi Modelli.

- **IBM SPSS Modeler - Nodi di modellazione.** Descrizioni di tutti i nodi utilizzati per creare modelli data mining. IBM SPSS Modeler offre numerosi metodi di modellazione ricavati dall'apprendimento automatico, dall'intelligenza artificiale e dalla statistica.
- **IBM SPSS Modeler Algorithms Guide.** Descrizione dei fondamenti di matematica per i metodi di modellazione utilizzati in IBM SPSS Modeler. Questa guida è disponibile solo in formato PDF.
- **IBM SPSS Modeler Applications Guide.** Gli esempi inclusi in questa guida forniscono indicazioni mirate e sintetiche su specifici metodi e tecniche di modellazione. Una versione in linea di questa guida è inoltre disponibile dal menu Aiuto. Per ulteriori informazioni, consultare l'argomento "Esempi di applicazioni" a pagina 5.
- **IBM SPSS Modeler - Guida per script e automazione.** Informazioni sulle modalità di automazione del sistema tramite script, incluse le proprietà che è possibile utilizzare per manipolare nodi e flussi.
- **IBM SPSS Modeler - Guida alla distribuzione.** Informazioni sull'esecuzione di flussi e scenari IBM SPSS Modeler come fasi dell'elaborazione di lavori in IBM SPSS Collaboration and Deployment Services Deployment Manager.
- **IBM SPSS Modeler - Guida per lo sviluppatore CLEF.** CLEF consente di integrare programmi di terze parti come routine di elaborazione dei dati o algoritmi di modellazione come nodi in IBM SPSS Modeler.
- **IBM SPSS Modeler - Guida al mining nel database.** Informazioni sulle modalità per utilizzare al meglio la potenza del database in uso al fine di ottenere prestazioni migliori ed estendere la gamma di funzionalità analitiche tramite algoritmi di terze parti.
- **IBM SPSS Modeler Server - Guida della performance e amministrazione.** Informazioni su come configurare e amministrare IBM SPSS Modeler Server.
- **IBM SPSS Modeler Administration Console - Guida per l'utente.** Informazioni sull'installazione e l'utilizzo dell'interfaccia utente della console per il monitoraggio e la configurazione di IBM SPSS Modeler Server. La console viene implementata come plug-in dell'applicazione Deployment Manager.
- **IBM SPSS Modeler - Guida CRISP-DM.** Guida passo a passo al data mining tramite la metodologia CRISP-DM con SPSS Modeler.
- **IBM SPSS Modeler Batch - Guida per l'utente.** Guida completa all'utilizzo di IBM SPSS Modeler in modalità batch, contenente dettagli per l'esecuzione della modalità batch e gli argomenti della riga di comando. Questa guida è disponibile solo in formato PDF.

Documentazione di SPSS Modeler Premium

La documentazione completa di SPSS Modeler Premium, escluse le istruzioni di installazione, è la seguente.

- **IBM SPSS Modeler Entity Analytics User Guide.** Contiene informazioni per l'utilizzo dell'analisi delle entità con SPSS Modeler; descrive l'installazione e la configurazione di repository, i nodi Entity Analytics e le attività amministrative.
- **IBM SPSS Modeler Social Network Analysis User Guide.** Guida che spiega come eseguire l'analisi dei social network con SPSS Modeler; comprende l'analisi di gruppo e l'analisi di diffusione.
- **SPSS Modeler Text Analytics - Guida per l'utente.** Contiene informazioni per l'utilizzo di analisi di testo con SPSS Modeler; descrive i nodi di text mining, il workbench interattivo, i modelli e altre risorse.
- **IBM SPSS Modeler Text Analytics Administration Console - Guida per l'utente.** Informazioni sull'installazione e l'utilizzo dell'interfaccia utente della console per il monitoraggio e la configurazione di IBM SPSS Modeler Server per l'utilizzo con SPSS Modeler Text Analytics. La console viene implementata come plug-in dell'applicazione Deployment Manager.

Esempi di applicazioni

Mentre gli strumenti per il data mining di SPSS Modeler consentono di risolvere un'ampia gamma di problemi a livello di business e organizzativo, gli esempi di applicazioni forniscono indicazioni mirate e sintetiche su specifici metodi e tecniche di modellazione. Gli insiemi di dati utilizzati negli esempi hanno dimensioni molto più limitate rispetto agli enormi archivi di dati gestiti da alcuni data miner, ma i concetti e i metodi coinvolti sono rapportabili alle applicazioni del mondo reale.

È possibile accedere agli esempi facendo clic su **Esempi di applicazioni** nel menu Aiuto di SPSS Modeler. I file di dati e i flussi di esempio sono installati nella cartella *Demos* nella directory di installazione del prodotto. Per ulteriori informazioni, consultare l'argomento "Cartella Demos".

Esempi di modellazione del database. Vedere gli esempi nella *IBM SPSS Modeler Guida al mining nel database*.

Esempi di script. Vedere gli esempi nella *IBM SPSS Modeler Guida per script e automazione*.

Cartella Demos

I file di dati e i flussi di esempio utilizzati negli esempi di applicazioni sono installati nella cartella *Demos* nella directory di installazione del prodotto. È possibile accedere a questa cartella anche dal gruppo di programmi IBM SPSS Modeler nel menu Start di Windows oppure facendo clic su *Demos* nell'elenco delle directory recenti nella finestra di dialogo Apri file.

Capitolo 2. Nodi origine

Osservazioni generali

I nodi origine consentono di importare i dati archiviati in numerosi formati, inclusi file flat, IBM SPSS Statistics (.sav), SAS, Microsoft Excel e database relazionali compatibili con ODBC. Tramite il nodo Input utente è inoltre possibile generare dati sintetici.

La palette Input contiene i seguenti nodi:



Il nodo Enterprise View crea una connessione a un IBM SPSS Collaboration and Deployment Services Repository, consentendo di leggere i dati di Enterprise View in un flusso e di assemblare un modello in uno scenario accessibile da altri utenti tramite il repository. Per ulteriori informazioni, consultare l'argomento "Nodo Enterprise View" a pagina 8.



Il nodo Vista dati può essere utilizzato per accedere alle origini dati definite nelle viste di dati analitici IBM SPSS Collaboration and Deployment Services. Una vista di dati analitici definisce un'interfaccia standard per l'accesso ai dati ed associa più origini dati fisiche a tale interfaccia. Per ulteriori informazioni, consultare l'argomento "Nodo Vista dati" a pagina 11.



Il nodo Database può essere utilizzato per importare dati da numerosi altri pacchetti che utilizzano ODBC (Open Database Connectivity), tra cui Microsoft SQL Server, DB2, Oracle e altri. Per ulteriori informazioni, consultare l'argomento "Nodo origine del database" a pagina 13.



Il nodo Testo variabile legge i dati dai file di testo a campi liberi — ovvero, file i cui record contengono un numero costante i campi ma un numero variabile di caratteri. Questo nodo può essere utilizzato per file con testo di intestazione a lunghezza fissa e alcuni tipi di annotazioni. Per ulteriori informazioni, consultare l'argomento "Nodo Testo variabile" a pagina 20.



Il nodo Testo fisso importa i dati dai file di testo a campi fissi — file i cui campi non sono delimitati e iniziano nella stessa posizione e sono di lunghezza fissa. Nel formato a campi fissi vengono in genere archiviati dati di versioni precedenti o generati dalla macchina. Per ulteriori informazioni, consultare l'argomento "Nodo Testo fisso" a pagina 22.



Il nodo File Statistics legge i dati dal formato file .sav o .zsav utilizzato da IBM SPSS Statistics, oltre ai file di cache salvati in IBM SPSS Modeler, che utilizzano lo stesso formato.



Il nodo IBM SPSS Data Collection importa i dati di indagine dai vari formati utilizzati dai software per ricerche di mercato conformemente al Modello dati di IBM SPSS Data Collection. Per utilizzare questo nodo, è necessario che sia installata IBM SPSS Data Collection Developer Library. Per ulteriori informazioni, consultare l'argomento "Nodo Data Collection" a pagina 25.



Il nodo origine IBM Cognos BI importa i dati dai database Cognos BI.



Il nodo origine IBM Cognos TM1 importa i dati dai database Cognos TM1.



Il nodo File SAS importa dati SAS in IBM SPSS Modeler. Per ulteriori informazioni, consultare l'argomento "Nodo origine SAS" a pagina 39.



Il nodo Excel importa dati da qualsiasi versione di Microsoft Excel. Non è richiesta alcuna sorgente dati ODBC. Per ulteriori informazioni, consultare l'argomento "Nodo origine Excel" a pagina 40.



Il nodo origine XML importa i dati in formato XML nel flusso. È possibile importare un singolo file oppure tutti i file in una directory. Come opzione, è possibile specificare un file schema da cui leggere la struttura XML.



Il nodo Genera simulazione fornisce un modo semplice per generare dati simulati, a partire da zero utilizzando le distribuzioni statistiche specificate dall'utente o automaticamente utilizzando le distribuzioni ottenute dall'esecuzione di un nodo Adattamento simulazione sui dati cronologici esistenti. Questo è utile quando si desidera valutare il risultato di un modello predittivo in presenza di incertezza negli input del modello.



Il nodo Input utente consente di creare facilmente dati sintetici — sia creandoli che modificando i dati esistenti. È utile, per esempio, quando si desidera creare un insieme di dati di test per la modellazione. Per ulteriori informazioni, consultare l'argomento "Nodo input utente" a pagina 55.

Per iniziare un flusso, aggiungere un nodo origine nell'area del flusso. Quindi fare doppio clic sul nodo per visualizzare la relativa finestra di dialogo. Le varie schede della finestra di dialogo consentono di leggere dati, visualizzare campi e valori e impostare numerose opzioni relative per esempio a filtri, tipi di dati, ruoli dei campi e controllo dei valori mancanti.

Nodo Enterprise View

Il nodo Enterprise View consente di creare e di mantenere una connessione tra una sessione di IBM SPSS Modeler e una Enterprise View in un IBM SPSS Collaboration and Deployment Services Repository condiviso. Ciò consente di leggere dati da una Enterprise View in un flusso IBM SPSS Modeler e di assemblare un modello IBM SPSS Modeler in uno scenario a cui possono accedere altri utenti del repository condiviso.

Uno **scenario** è un file contenente un flusso di IBM SPSS Modeler con determinati nodi, modelli e proprietà aggiuntive che ne consentono la distribuzione in IBM SPSS Collaboration and Deployment Services Repository ai fini del calcolo del punteggio o per l'aggiornamento automatico di modelli. L'uso

di nodi Enterprise View con scenari garantisce che, in una situazione multi-utente, tutti gli utenti lavorino sugli stessi dati. Una **connessione** è un collegamento da una sessione di IBM SPSS Modeler a una Enterprise View nel IBM SPSS Collaboration and Deployment Services Repository.

L'**Enterprise View** è l'insieme completo di dati appartenenti a un'organizzazione, indipendentemente dalla posizione fisica in cui risiedono i dati. Ogni connessione è costituita da una selezione specifica di un'unica **Visualizzazione applicazione** (sottoinsieme di Enterprise View su misura per una particolare applicazione), una **Definizione dei provider di dati** (DPD--collega le tabelle e le colonne della Visualizzazione applicazione con la sorgenti dati fisica) e un **ambiente** (identifica quali specifiche colonne devono essere associate ai segmenti di business definiti). Le definizioni Enterprise View, Visualizzazioni applicazione e DPD vengono archiviate nel repository, sebbene i dati effettivi risiedano in uno o più database o in altre sorgenti esterne.

Quando la connessione è stata stabilita, è necessario specificare una **tabella** di Visualizzazione applicazione con cui lavorare in IBM SPSS Modeler. In una Visualizzazione applicazione, una tabella è una visualizzazione logica costituita da alcune o tutte le colonne di una o più tabelle fisiche in uno o più database. Pertanto il nodo Enterprise View consente di visualizzare in IBM SPSS Modeler i record provenienti da più tabelle di database come un'unica tabella.

Requisiti

- Per utilizzare il nodo Enterprise View, è necessario dapprima installare e configurare nel sito un IBM SPSS Collaboration and Deployment Services Repository, con una Enterprise View, le Visualizzazioni applicazione e le DPD già definite.

Nota: per l'accesso ad un repository IBM SPSS Collaboration and Deployment Services è richiesta una licenza separata. Per ulteriori informazioni, consultare <http://www.ibm.com/software/analytics/spss/products/deployment/cds/>

- È inoltre necessario che IBM SPSS Collaboration and Deployment Services Enterprise View Driver sia installato su ogni computer utilizzato per modificare o eseguire il flusso. Per Windows, installare semplicemente il driver sul computer su cui è installato IBM SPSS Modeler o IBM SPSS Modeler Server. Non sono necessarie ulteriori configurazioni del driver. In UNIX, è necessario aggiungere allo script di avvio un riferimento allo script *pev.sh*. Rivolgersi all'amministratore locale per i dettagli relativi all'installazione di IBM SPSS Collaboration and Deployment Services Enterprise View Driver.
- Una Definizione dei provider di dati (DPD) viene definita rispetto a una determinata sorgente dati ODBC. Per utilizzare una DPD da IBM SPSS Modeler, è necessario che sull'host di IBM SPSS Modeler sia stata definita una sorgente dati ODBC con lo stesso nome e che si connette allo stesso archivio dati di quella a cui fa riferimento la Definizione dei provider di dati.

Impostazione delle opzioni relative al nodo Enterprise View

Le opzioni della scheda Dati della finestra di dialogo Enterprise View possono essere utilizzate per:

- selezionare la connessione a un repository esistente
- modificare la connessione a un repository esistente
- creare una nuova connessione a un repository
- selezionare una tabella di Visualizzazione applicazione

Per ulteriori informazioni su come utilizzare i repository, consultare *IBM SPSS Collaboration and Deployment Services Administrator's Guide*.

Connessione. L'elenco a discesa contiene le opzioni per la selezione della connessione a un repository esistente, la modifica di una connessione esistente o l'aggiunta di una connessione. Se si è già connessi a un repository tramite IBM SPSS Modeler, scegliendo l'opzione **Aggiungi/modifica una connessione** viene visualizzata la finestra di dialogo Connessioni Enterprise View che consente di definire o modificare i dettagli necessari per la connessione corrente. Se non si è connessi, questa opzione visualizza la finestra di dialogo Accesso del repository.

Per informazioni relative all'accesso al repository, consultare *IBM SPSS Modeler - Guida per l'utente*.

Dopo aver stabilito una connessione a un repository, tale connessione rimane attiva finché non si esce da IBM SPSS Modeler. È possibile condividere una connessione con altri nodi all'interno dello stesso stream, ma è necessario creare una nuova connessione per ogni nuovo stream.

Dopo un accesso riuscito, viene visualizzata la finestra di dialogo Connessioni Enterprise View.

Nome tabella. All'inizio questo campo è vuoto e non è possibile compilarlo fino a quando non si crea una connessione. Se si conosce il nome della tabella di Visualizzazione applicazione a cui si desidera accedere, digitarlo nel campo Nome tabella. In alternativa, fare clic sul pulsante **Seleziona** per aprire una finestra di dialogo contenente l'elenco delle tabelle di Visualizzazione applicazione disponibili.

Connessioni Enterprise View

Questa finestra di dialogo consente di definire o di modificare i dati necessari per la connessione al repository. È possibile specificare i seguenti dati:

- Visualizzazione applicazione e versione
- Ambiente
- Definizione dei provider di dati (DPD)
- Descrizione della connessione

Connessioni. Elenca le connessioni ai repository esistenti.

- **Aggiungi una nuova connessione.** Visualizza la finestra di dialogo Recupera oggetto che consente di cercare e selezionare una Visualizzazione applicazione nel repository.
- **Copia la connessione selezionata.** Crea una copia della connessione selezionata, evitando all'utente di dover nuovamente cercare la stessa Visualizzazione applicazione.
- **Elimina la connessione selezionata.** Elimina dall'elenco la connessione selezionata.

Dettagli di connessione. Per la connessione attualmente selezionata nel riquadro Connessioni, visualizza la Visualizzazione applicazione, l'etichetta della versione, l'ambiente, la Definizione dei provider di dati e un testo descrittivo.

- **Visualizzazione applicazione.** In questo elenco a discesa viene visualizzata l'eventuale Visualizzazione applicazione selezionata. Se nella sessione corrente sono state stabilite connessioni ad altre visualizzazioni applicazione, nell'elenco a discesa saranno visualizzate anche queste. Fare clic sul pulsante Visualizza adiacente per cercare altre Visualizzazioni applicazione nel repository.
- **Etichetta versione.** L'elenco a discesa mostra tutte le etichette delle versioni definite per la Visualizzazione applicazione specificata. Le etichette delle versioni facilitano l'individuazione di versioni specifiche di oggetti nel repository. Per esempio, potrebbero esistere due versioni di una determinata Visualizzazione applicazione. Utilizzando le etichette si potrebbe impostare l'etichetta **TEST** per la versione utilizzata nell'ambiente di sviluppo e l'etichetta **PRODUZIONE** per la versione utilizzata nell'ambiente di produzione. Selezionare un'etichetta appropriata.

Nota: le etichette non devono contenere il carattere "["; in caso contrario, il nome della tabella non verrà visualizzato nella scheda Dati della finestra di dialogo Enterprise View.

- **Ambiente.** L'elenco a discesa mostra tutti gli ambienti validi. L'impostazione Ambiente determina quali DPD sono disponibili. Specifica pertanto quali colonne specifiche devono essere associate ai segmenti di business definiti. Per esempio, se si seleziona **Analitico**, vengono restituite solo le colonne di Visualizzazione applicazione definite come **Analitiche**. L'ambiente di default è **Analitico**; è possibile anche scegliere l'opzione **Operativo**.
- **Provider di dati.** Nell'elenco a discesa sono visualizzati fino a dieci nomi di Definizioni dei provider di dati per la Visualizzazione applicazione selezionata. Vengono visualizzate solo le Definizioni dei provider di dati che fanno riferimento alla Visualizzazione applicazione selezionata. Fare clic sul

pulsante Visualizza adiacente per visualizzare il nome e il percorso di tutte le Definizioni dei provider di dati relative alla Visualizzazione applicazione corrente.

- **Descrizione.** Testo descrittivo relativo alla connessione con il repository. Il testo sarà utilizzato per il nome della connessione: se si fa clic su **OK**, il testo viene visualizzato nell'elenco a discesa Connessione e nella barra del titolo della finestra di dialogo Enterprise View, nonché come etichetta del nodo Enterprise View nell'area.

Scelta della Definizione dei provider di dati

La finestra di dialogo Seleziona provider di dati mostra il nome e il percorso di tutte le Definizioni dei provider di dati relative alla Visualizzazione applicazione corrente.

Le Visualizzazioni applicazione possono avere più Definizioni dei provider di dati per supportare varie fasi di un progetto. Per esempio, i dati cronologici utilizzati per generare un modello potrebbero provenire da un database, mentre i dati operativi potrebbero risiedere in un altro.

Una Definizione dei provider di dati (DPD) viene definita rispetto a una determinata sorgente dati ODBC. Per utilizzare una DPD da IBM SPSS Modeler, è necessario che sull'host di IBM SPSS Modeler sia stata definita una sorgente dati ODBC con lo stesso nome e che si connette allo stesso archivio dati di quella a cui fa riferimento la Definizione dei provider di dati.

Per scegliere una Definizione dei provider di dati con cui lavorare, selezionarne il nome nell'elenco, quindi fare clic su **OK**.

Scelta della tabella

La finestra di dialogo Seleziona tabella elenca tutte le tabelle a cui viene fatto riferimento nella Visualizzazione applicazione corrente. La finestra di dialogo è vuota se non è stata effettuata alcuna connessione a un IBM SPSS Collaboration and Deployment Services Repository.

Per scegliere una tabella con cui lavorare, selezionarne il nome nell'elenco, quindi fare clic su **OK**.

Nodo Vista dati

Utilizzare il nodo Vista dati per includere i dati definiti in una vista dati analitici IBM SPSS Collaboration and Deployment Services nel flusso. Una vista di dati analitici definisce una struttura per l'accesso ai dati che descrivono le entità utilizzate nei modelli predittivi e nelle regole di business. La vista associa la struttura dei dati alle origini fisiche dei dati per l'analisi.

L'analisi predittiva richiede che i dati siano organizzati in tabelle dove ogni riga corrisponde ad un'entità per la quale vengono effettuate le previsioni. Ogni colonna di una tabella rappresenta un attributo misurabile dell'entità. Alcuni attributi possono essere ricavati dall'aggregazione dei valori di un altro attributo. Ad esempio, le righe di una tabella potrebbero rappresentare i clienti con le colonne che corrispondono al nome cliente, al sesso, al CAP e al numero di volte in cui il cliente ha effettuato un acquisto superiore a \$500 nell'ultimo anno. L'ultima colonna viene ricavata dalla cronologia degli ordini del cliente, che in genere viene memorizzata in una o più tabelle correlate.

Il processo di analisi predittiva implica l'utilizzo di serie diverse di dati nel ciclo di vita di un modello. Durante lo sviluppo iniziale di un modello predittivo, vengono utilizzati dati cronologici che spesso hanno risultati noti per l'evento in fase di previsione. Per valutare l'efficacia e l'accuratezza del modello, viene convalidato un modello candidato rispetto a dati diversi. Dopo aver convalidato il modello, viene distribuito per l'utilizzo in produzione per generare punteggi per più entità in un processo batch o per singole entità in un processo in tempo reale. Se si combina il modello con le regole di business in un processo di gestione decisioni, vengono utilizzati i dati simulati per convalidare i risultati della combinazione. Tuttavia, sebbene i dati utilizzati differiscono nelle fasi del processo di sviluppo del modello, ogni dataset deve fornire la stessa serie di attributi per il modello. La serie di attributi resta costante; i record dei dati analizzati cambiano.

Una vista di dati analitici è composta dai seguenti componenti che gestiscono le esigenze specializzate dell'analisi predittiva:

- Uno schema vista dati o modello dati che definisce un'interfaccia logica per accedere ai dati come un insieme di attributi organizzati in tabelle correlate. Gli attributi nel modello possono essere ricavati da altri attributi.
- Uno o più piani di accesso ai dati forniscono gli attributi del modello dati con valori fisici. Per controllare i dati disponibili per il modello dati specificare quale piano di accesso dati è attivo per una particolare applicazione.

Importante: Per utilizzare il nodo Vista dati, è necessario prima che sia installato e configurato un IBM SPSS Collaboration and Deployment Services Repository nel sito. La vista di dati analitici a cui fa riferimento il nodo in genere viene creata e memorizzata nel repository tramite IBM SPSS Collaboration and Deployment Services Deployment Manager.

Impostazione delle opzioni relative al nodo Vista dati

Utilizzare le opzioni presenti nella scheda **Dati** della finestra di dialogo del nodo Vista dati per specificare le impostazioni dei dati per una vista di dati analitici selezionata dal IBM SPSS Collaboration and Deployment Services Repository.

Vista di dati analitici. Fare clic sul pulsante con i puntini sospensivi (...) per selezionare una vista di dati analitici. Se al momento non si è connessi ad un server di repository, specificare l'URL per il server nella finestra di dialogo Repository: Server, fare clic su **OK** e specificare le credenziali della connessione nella finestra di dialogo Repository: Credenziali. Per ulteriori informazioni su come accedere al repository e richiamare gli oggetti, consultare IBM SPSS Modeler - Guida per l'utente.

Nome tabella. Selezionare una tabella dal modello dati nella vista di dati analitici. Ogni tabella nel modello dati rappresenta un concetto o un'entità implicati nel processo analitico predittivo. I campi per le tabelle corrispondono agli attributi delle entità rappresentati dalle tabelle. Ad esempio, se si stanno analizzando gli ordini dei clienti, il modello dati potrebbe includere una tabella per i clienti e una tabella per gli ordini. La tabella dei clienti potrebbe contenere attributi per l'identificativo del cliente, l'età, il sesso, lo stato civile e il paese di residenza. La tabella degli ordini potrebbe contenere attributi per l'identificativo dell'ordine, il numero di elementi nell'ordine, il costo totale e l'identificativo del cliente che ha effettuato l'ordine. L'attributo dell'identificativo del cliente potrebbe essere utilizzato per associare i clienti nella tabella dei clienti ai relativi ordini nella tabella degli ordini.

Piano di accesso ai dati. Selezionare un piano di accesso ai dati dalla vista di dati analitici. Un piano di accesso ai dati associa le tabelle dei modelli di dati in una vista di dati analitici alle origini dati fisiche. Una vista di dati analitici in genere contiene più piani di accesso ai dati. Quando si modifica il piano di accesso ai dati in uso, si modificano i dati utilizzati dal flusso. Ad esempio, se la vista di dati analitici contiene un piano di accesso ai dati per l'addestramento di un modello e un piano di accesso ai dati per la verifica di un modello, è possibile passare dai dati di addestramento a quelli di verifica cambiando il piano di accesso ai dati in uso.

Attributi facoltativi. Se un particolare attributo non è richiesto dall'applicazione che utilizza la vista di dati analitici, è possibile contrassegnare l'attributo come facoltativo. Diversamente dagli attributi richiesti, gli attributi facoltativi possono includere i valori null. Potrebbe essere necessario adattare l'applicazione includendo la gestione di valori null per gli attributi facoltativi. Ad esempio, quando si richiama una regola di business creata in IBM Operational Decision Manager, IBM Analytical Decision Management esegue una query del servizio di regole per determinare quali input sono richiesti. Se il record di cui si deve calcolare il punteggio contiene un valore null per uno dei campi richiesti del servizio di regole, la regola non viene richiamata e i campi di output della regola vengono popolati con valori di default. Se un campo facoltativo contiene un valore null, la regola viene richiamata. La regola può verificare la presenza di valori null per controllare l'elaborazione.

Per specificare gli attributi come facoltativi, fare clic su **Attributi facoltativi** e selezionare gli attributi facoltativi.

Includi dati XML nel campo. Selezionare questa opzione per creare un campo contenente i dati XML del modello di oggetto eseguibile per ciascuna riga di dati. Queste informazioni sono richieste se i dati verranno utilizzati con IBM Operational Decision Manager. Specificare il nome di questo nuovo campo.

Nodo origine del database

Il nodo origine del database può essere utilizzato per importare dati da numerosi altri pacchetti che utilizzano ODBC (Open Database Connectivity), tra cui Microsoft SQL Server, DB2, Oracle e altri.

Per leggere o scrivere su un database, occorre che un'origine dati ODBC sia installata e configurata per il database in questione, con le relative autorizzazioni di lettura e scrittura. IBM SPSS Data Access Pack include una serie di driver ODBC che è possibile utilizzare per questo scopo, disponibili nel DVD IBM SPSS Data Access Pack oppure dal sito di download. Per domande su come creare o impostare le autorizzazioni per le origini dati ODBC, contattare l'amministratore di database.

Driver ODBC supportati

Per le ultime informazioni sui database e i driver ODBC che sono supportati e testati per l'uso con IBM SPSS Modeler 16, vedere le matrici di compatibilità del prodotto sul sito aziendale dedicato all'assistenza clienti (<http://www.ibm.com/support>).

Dove installare i driver

I driver ODBC devono essere installati e configurati su ogni computer utilizzato.

- Se IBM SPSS Modeler viene eseguito in modalità locale (standalone), i driver devono essere installati sul computer locale.
- Se IBM SPSS Modeler viene eseguito in modalità distribuita su un IBM SPSS Modeler Server remoto, i driver ODBC devono essere installati nel computer in cui è installato IBM SPSS Modeler Server. Per IBM SPSS Modeler Server su sistemi UNIX, vedere anche "Configurazione dei driver ODBC in sistemi UNIX" più avanti in questa sezione.
- Se occorre accedere alle stesse origini dati sia da IBM SPSS Modeler che da IBM SPSS Modeler Server, i driver ODBC devono essere installati su entrambi i computer.
- Se IBM SPSS Modeler è in esecuzione su Terminal Services, i driver ODBC devono essere installati sul server Terminal Services in cui è stato installato IBM SPSS Modeler.

Importante: se si utilizza IBM SPSS Modeler Server su UNIX per accedere ad un database Teradata e si utilizza il driver ODBC fornito da Teradata, è necessario utilizzare il gestore driver ODBC installato con esso. (**Nota:** non è necessario apportare queste modifiche se si utilizza il driver SDAP di Teradata). Per poter apportare questa modifica a IBM SPSS Modeler Server, specificare un valore per ODBC_DRIVER_MANAGER_PATH all'inizio dello script modelersrv.sh dove indicato dai commenti. Questa variabile di ambiente deve essere impostata sul percorso del gestore dei driver ODBC, che viene fornito con il driver ODBC Teradata (/usr/odbc/lib in un'installazione predefinita del driver ODBC Teradata). È necessario riavviare IBM SPSS Modeler Server per applicare la modifica. Per i dettagli relativi alle piattaforme IBM SPSS Modeler Server che offrono supporto all'accesso Teradata, e la versione di driver ODBC Teradata supportata, vedere il sito aziendale dedicato al supporto tecnico all'indirizzo <http://www.ibm.com/support>.

Configurazione dei driver ODBC in sistemi UNIX

Di default, il gestore dei driver DataDirect non è configurato per IBM SPSS Modeler Server nei sistemi UNIX. Per configurare UNIX per il caricamento del gestore dei driver DataDirect, immettere i comandi seguenti:

```
cd modeler_server_install_directory/bin
rm -f libspssodbc.so
ln -s libspssodbc_datadirect.so libspssodbc.so
```

In questo modo si rimuove il collegamento di default e si crea un collegamento al gestore dei driver DataDirect.

Utilizzare i passaggi seguenti per accedere ai dati di un database:

1. Installare un driver ODBC e configurare una sorgente dati per il database che si desidera utilizzare.
2. Nella finestra di dialogo del nodo Database eseguire la connessione a un database utilizzando la modalità Tabella o Query SQL.
3. Selezionare una tabella del database.
4. Tramite le schede della finestra di dialogo del nodo Database è possibile modificare i tipi di utilizzo e filtrare i campi dei dati.

Questi passaggi verranno descritti in dettaglio negli argomenti seguenti.

Impostazione delle opzioni relative al nodo Database

È possibile utilizzare le opzioni nella scheda Dati della finestra di dialogo del nodo origine del database per ottenere l'accesso a un database e leggere i dati dalla tabella selezionata.

Modalità. Selezionare **Tabella** per effettuare la connessione ad una tabella utilizzando i controlli della finestra di dialogo.

Selezionare **Query SQL** per eseguire query nel database selezionato nel campo sottostante utilizzando SQL. Per ulteriori informazioni, consultare l'argomento "Invio di query al database" a pagina 18.

Sorgente dati. Per le modalità Tabella e Query SQL, è possibile immettere un nome nel campo Sorgente dati oppure selezionare **Aggiungi nuova connessione database** dall'elenco a discesa.

Le opzioni seguenti vengono utilizzate per eseguire la connessione a un database e selezionare una tabella utilizzando la finestra di dialogo:

Nome tabella. Se si conosce il nome della tabella a cui si desidera accedere, immetterlo nel campo Nome tabella. In alternativa, fare clic sul pulsante **Seleziona** per aprire una finestra di dialogo contenente l'elenco delle tabelle disponibili.

Nomi di tabelle e colonne fra virgolette. Specificare se si desidera che i nomi delle tabelle e delle colonne vengano racchiusi tra virgolette quando le query vengono inviate al database (per esempio nel caso in cui tali nomi contengano spazi o punteggiatura).

- L'opzione **Se necessario** consente di racchiudere nomi di tabelle e campi tra virgolette *solo* se contengono caratteri non standard. I caratteri non standard sono rappresentati da caratteri ASCII, spazi e qualsiasi carattere non alfanumerico diverso dal punto (.).
- Selezionare **Mai** se si desidera che i nomi di campi e tabelle non vengano *mai* racchiusi tra virgolette.
- Selezionare **Sempre** se si desidera che *tutti* i nomi di campi e tabelle vengano racchiusi tra virgolette.

Elimina spazi all'inizio e alla fine. Selezionare le opzioni per l'eliminazione degli spazi iniziali e finali nelle stringhe.

Nota: i confronti tra stringhe che utilizzano e non utilizzano il push back SQL potrebbero generare risultati diversi in presenza di spazi finali.

Lettura di stringhe vuote da Oracle. Durante la lettura o la scrittura in un database Oracle, tenere presente che, a differenza di IBM SPSS Modeler e a differenza della maggior parte degli altri database,

Oracle considera ed archivia i valori di stringa vuoti come valori null. Questo significa che gli stessi dati estratti da un database Oracle potrebbero comportarsi in modo diverso rispetto a quando vengono estratti da un file o da un altro database e i dati potrebbero restituire risultati diversi.

Aggiunta di una connessione a un database

Per aprire un database è innanzitutto necessario selezionare la sorgente dati a cui si desidera eseguire la connessione. Nella scheda Dati, selezionare **Aggiungi nuova connessione database** dall'elenco a discesa Sorgente dati.

In questo modo viene aperta la finestra di dialogo Connessioni database.

Nota: in alternativa, per visualizzare questa finestra di dialogo, dal menu principale selezionare:

Strumenti > Database...

Sorgenti dati. Elenca le origini dati disponibili. Scorrere completamente l'elenco se non viene visualizzato il database desiderato. Dopo aver selezionato una sorgente dati e immesso eventuali password, fare clic su **Connetti**. Fare clic su **Aggiorna** per aggiornare l'elenco.

Nome utente. Se l'origine dati è protetta da password, immettere il proprio nome utente.

Password. Se l'origine dati è protetta da password, immettere la propria password.

Connessioni. Mostra i database attualmente connessi.

- **Default.** È possibile scegliere una connessione come valore di default. In questo caso, i nodi Database origine o di esportazione avranno come sorgente dati questa connessione predefinita, anche se l'impostazione si può modificare.
- **Salva.** Se lo si desidera, selezionare una o più connessioni da visualizzare nuovamente nelle sessioni successive.
- **Origine dati.** Le stringhe di connessione per i database connessi in quel momento.
- **Preimpostati.** Indica (con un carattere *) se sono stati specificati dei valori preimpostati per la connessione al database. Per specificare valori preimpostati, fare clic su questa colonna nella riga corrispondente alla connessione al database e scegliere Specifica dall'elenco. Per ulteriori informazioni, consultare l'argomento "Specificazione di valori preimpostati per una connessione a un database".

Per rimuovere una connessione esistente, selezionarla dall'elenco e fare clic su **Rimuovi**.

Una volta effettuate tutte le selezioni, fare clic su **OK**.

- Se IBM SPSS Modeler viene eseguito in modalità locale (standalone), i driver devono essere installati sul computer locale.
- Se IBM SPSS Modeler viene eseguito in modalità distribuita su un IBM SPSS Modeler Server remoto, i driver ODBC devono essere installati nel computer in cui è installato IBM SPSS Modeler Server. Per IBM SPSS Modeler Server su sistemi UNIX, vedere anche "Configurazione dei driver ODBC in sistemi UNIX" più avanti in questa sezione.
- Se occorre accedere alle stesse origini dati sia da IBM SPSS Modeler che da IBM SPSS Modeler Server, i driver ODBC devono essere installati su entrambi i computer.
- Se IBM SPSS Modeler è in esecuzione su Terminal Services, i driver ODBC devono essere installati sul server Terminal Services in cui è stato installato IBM SPSS Modeler.

Specificazione di valori preimpostati per una connessione a un database

Per alcuni database, è possibile specificare numerose impostazioni di default per la connessione al database. Le impostazioni si applicano tutte all'esportazione di database.

Questa funzione è supportata dai tipi di database elencati di seguito.

- IBM InfoSphere Warehouse eseguito su DB2 9.1 o successiva. Per ulteriori informazioni, consultare l'argomento "Impostazioni per IBM DB2 InfoSphere Warehouse".
- SQL Server 2008 o le successive versioni Enterprise e Developer. Per ulteriori informazioni, consultare l'argomento "Impostazioni per SQL Server".
- Oracle 10g e 11gR1 o le successive versioni Enterprise o Personal. Per ulteriori informazioni, consultare l'argomento "Impostazioni per Oracle".
- IBM Netezza, IBM DB2 for z/OS e Teradata effettuano la connessione ad un database o uno schema in modo simile. Per ulteriori informazioni, consultare l'argomento "Impostazioni per IBM Netezza, IBM DB2 for z/OS, IBM DB2 LUW e Teradata" a pagina 17.

Se si è connessi ad un database o uno schema che non supporta tale funzione, viene visualizzato il messaggio **Non è possibile configurare alcuna preimpostazione per questa connessione database.**

Impostazioni per IBM DB2 InfoSphere Warehouse

Queste impostazioni vengono visualizzate per IBM InfoSphere Warehouse eseguito su DB2 9.1 o successive.

Spazio di tabella. Lo spazio di tabella da utilizzare per l'esportazione. Gli amministratori di database possono creare o configurare spazi di tabella partizionati. Si consiglia di selezionare uno di questi spazi di tabella (anziché quello di default) da utilizzare per le esportazioni in database.

Utilizza compressione. Se questa opzione è selezionata, vengono create le tabelle per l'esportazione con la compressione (ad esempio, l'equivalente di CREATE TABLE MYTABLE(...) COMPRESS YES; in SQL).

Non registrare gli aggiornamenti. Se questa opzione è selezionata, non viene eseguita la registrazione durante la creazione delle tabelle e l'inserimento di dati (l'equivalente di CREATE TABLE MYTABLE(...) NOT LOGGED INITIALLY; in SQL).

Impostazioni per SQL Server

Queste impostazioni vengono visualizzate per SQL Server 2008 o versioni Enterprise e Developer successive.

Utilizza compressione. Se è selezionata, questa opzione utilizza la compressione nella creazione delle tabelle per l'esportazione.

Compressione per. Scegliere il livello di compressione.

- **Riga.** Abilita la compressione a livello di riga (ad esempio, l'equivalente di CREATE TABLE MYTABLE(...) WITH (DATA_COMPRESSION = ROW); in SQL).
- **Pagina.** Abilita la compressione a livello di pagina (ad esempio, CREATE TABLE MYTABLE(...) WITH (DATA_COMPRESSION = PAGE); in SQL).

Impostazioni per Oracle

Impostazioni di Oracle 10g

Queste impostazioni vengono visualizzate per Oracle 10g versione Enterprise o Personal.

Utilizza compressione. Se è selezionata, questa opzione utilizza la compressione nella creazione delle tabelle per l'esportazione. Per questa versione del database, è disponibile solo la compressione di base (ad esempio, CREATE TABLE MYTABLE(...) COMPRESS; in SQL).

Impostazioni di Oracle 11gR1

Queste impostazioni vengono visualizzate per Oracle 11g R1 versione Enterprise o Personal.

Utilizza compressione. Se è selezionata, questa opzione utilizza la compressione nella creazione delle tabelle per l'esportazione.

Compressione per. Scegliere il livello di compressione.

- **Default.** Abilita la compressione di default (ad esempio, CREATE TABLE MYTABLE(...) COMPRESS; in SQL). In questo caso, svolge la stessa funzione dell'opzione **Operazioni caricamento diretto**.
- **Operazioni caricamento diretto.** Abilita la compressione solo per le operazioni di inserimento di massa (percorso diretto) (ad esempio, CREATE TABLE MYTABLE(...)COMPRESS FOR DIRECT_LOAD OPERATIONS; in SQL).
- **Tutte le operazioni.** Abilita la compressione per tutte le operazioni (ad esempio, CREATE TABLE MYTABLE(...)COMPRESS FOR ALL OPERATIONS; in SQL).

Impostazioni di Oracle 11gR2 - Opzione di base

Queste impostazioni vengono visualizzate per Oracle 11g R2 versione Enterprise o Personal con opzione di base.

Utilizza compressione. Se è selezionata, questa opzione utilizza la compressione nella creazione delle tabelle per l'esportazione.

Compressione per. Scegliere il livello di compressione.

- **Default.** Abilita la compressione di default (ad esempio, CREATE TABLE MYTABLE(...) COMPRESS; in SQL). In questo caso, svolge la stessa funzione dell'opzione **Di base**.
- **Di base.** Abilita la compressione di base (ad esempio, CREATE TABLE MYTABLE(...) COMPRESS BASIC; in SQL).

Impostazioni di Oracle 11gR2 - Opzione avanzata

Queste impostazioni vengono visualizzate per Oracle 11g R2 versione Enterprise o Personal con opzione avanzata.

Utilizza compressione. Se è selezionata, questa opzione utilizza la compressione nella creazione delle tabelle per l'esportazione.

Compressione per. Scegliere il livello di compressione.

- **Default.** Abilita la compressione di default (ad esempio, CREATE TABLE MYTABLE(...) COMPRESS; in SQL). In questo caso, svolge la stessa funzione dell'opzione **Di base**.
- **Di base.** Abilita la compressione di base (ad esempio, CREATE TABLE MYTABLE(...) COMPRESS BASIC; in SQL).
- **OLTP.** Abilita la compressione OLTP (ad esempio, CREATE TABLE MYTABLE(...)COMPRESS FOR OLTP; in SQL).
- **Min/Max query.** (Solo server Exadata) Abilita la compressione colonnare ibrida per la query (ad esempio, CREATE TABLE MYTABLE(...)COMPRESS FOR QUERY LOW; o CREATE TABLE MYTABLE(...)COMPRESS FOR QUERY HIGH; in SQL). La compressione per la query è utile negli ambienti data warehousing; HIGH fornisce un rapporto di compressione maggiore rispetto a LOW.
- **Min/Max archivio.** (Solo server Exadata) Abilita la compressione colonnare ibrida per l'archivio (ad esempio, CREATE TABLE MYTABLE(...)COMPRESS FOR ARCHIVE LOW; o CREATE TABLE MYTABLE(...)COMPRESS FOR ARCHIVE HIGH; in SQL). La compressione per l'archivio è utile per la compressione dei dati che verranno archiviati per lunghi periodi di tempo; HIGH fornisce un rapporto di compressione maggiore rispetto a LOW.

Impostazioni per IBM Netezza, IBM DB2 for z/OS, IBM DB2 LUW e Teradata

Quando si specificano preimpostazioni per IBM Netezza, IBM DB2 for z/OS, IBM DB2 LUW o Teradata all'utente viene richiesto di selezionare quanto segue:

Utilizza schema / database adattatore per calcolo del punteggio server. Se questa opzione è selezionata, viene abilitata l'opzione **Schema/database adattatore per calcolo del punteggio server**.

Schema/database adattatore per calcolo del punteggio server. Dall'elenco a discesa, selezionare la connessione richiesta.

Selezione di una tabella di database

Dopo aver eseguito la connessione a una sorgente dati, è possibile scegliere di importare campi da una tabella o una vista specifica. Dalla scheda Dati della finestra di dialogo Database, è possibile immettere il nome di una tabella nel campo Nome tabella oppure fare clic su **Seleziona** per visualizzare la finestra di dialogo Seleziona tabella/vista che elenca le tabelle e le viste disponibili.

Mostra proprietario tabelle. Selezionare se un'origine dati richiede che venga specificato il nome del proprietario della tabella prima che sia possibile accedere alla tabella. Deselezionare questa opzione per le sorgenti dati che non includono questo requisito.

Nota: in genere, per i database SAS e Oracle è necessario specificare il proprietario della tabella.

Tabelle/viste. Selezionare la tabella o la vista da importare.

Mostra. Elenca le colonne nell'origine dati a cui è attualmente stabilita una connessione. Fare clic su una delle opzioni seguenti per personalizzare la visualizzazione delle tabelle disponibili:

- Selezionare **Tabelle utente** per visualizzare le normali tabelle di database create dagli utenti del database.
- Selezionare **Tabelle di sistema** per visualizzare le tabelle di database il cui proprietario è il sistema, per esempio le tabelle che forniscono le informazioni relative al database, come i dettagli sugli indici. Questa opzione può essere usata per visualizzare le schede utilizzate in database di Excel (si noti che è disponibile anche un nodo origine Excel separato. Per ulteriori informazioni, consultare l'argomento "Nodo origine Excel" a pagina 40.
- Selezionare **Viste** per visualizzare tabelle virtuali risultanti da una query su una o più tabelle ordinarie.
- Selezionare **Sinonimi** per visualizzare i sinonimi creati nel database per tutte le tabelle esistenti.

Filtri Nome/Proprietario. Tali campo consentono di filtrare l'elenco delle tabelle visualizzate in base al nome o al proprietario. Per esempio, digitare SYS per visualizzare l'elenco delle tabelle con il proprietario specificato. Per le ricerche tramite i caratteri jolly, è possibile utilizzare il carattere di sottolineatura (_) per rappresentare qualsiasi carattere singolo e il segno di percentuale (%) per rappresentare una sequenza di zero o più caratteri.

Imposta come default. Salva le impostazioni correnti come impostazioni di default per l'utente corrente. Le impostazioni verranno ripristinate successivamente quando un utente apre una nuova finestra di selezione delle tabelle *per lo stesso nome di sorgente dati e login utente*.

Invio di query al database

Dopo aver eseguito una connessione a una sorgente dati è possibile scegliere di importare campi utilizzando le query SQL. Nella finestra di dialogo principale selezionare **Query SQL** come modalità di connessione. In questo modo verrà aggiunta una finestra dell'editor di query nella finestra di dialogo. Grazie all'editor di query è possibile creare o caricare una o più query SQL il cui insieme di risultati verrà letto nel flusso di dati.

Se si specificano più query SQL, separarle con un punto e virgola (;) e assicurarsi che non ci sia più di un'istruzione SELECT.

Per chiudere la finestra dell'editor di query, selezionare **Tabella** come modalità di connessione.

È possibile inserire parametri di stream SPSS Modeler (un tipo di variabile definita dall'utente) nella query SQL. Per ulteriori informazioni, consultare l'argomento "Utilizzo dei parametri dei flussi in una query SQL".

Carica query. Fare clic per visualizzare il browser di file, che è possibile utilizzare per caricare una query salvata precedentemente.

Salva query. Fare clic per visualizzare la finestra di dialogo Salva query, che è possibile utilizzare per salvare la query corrente.

Importa default. Fare clic per importare un'istruzione SQL SELECT di esempio creata automaticamente utilizzando la tabella e le colonne selezionate nella finestra di dialogo.

Cancella. Cancella il contenuto dell'area di lavoro. Utilizzare questa opzione se si desidera cominciare da capo.

Suddividi testo. L'opzione predefinita **Mai** indica che la query verrà inviata al database come intera. Alternativamente, è possibile scegliere **Se necessario** il che indica che SPSS Modeler tenta di analizzare la query e di identificare se sono presenti istruzioni SQL che dovrebbero essere inviate al database una dopo l'altra.

Utilizzo dei parametri dei flussi in una query SQL

Durante la scrittura di una query SQL per importare i campi, è possibile inserire i parametri del flusso SPSS Modeler definiti in precedenza. Sono supportati tutti i tipi di parametri del flusso.

La tabella seguente descrive come vengono interpretati alcuni parametri dei flussi nella query SQL.

Tabella 1. Esempi di parametri dei flussi.

Nome del parametro del flusso (esempio)	Archiviazione	Valore del parametro del flusso	Interpretato come
PString	Stringa	ss	'ss'
PInt	Intero	5	5
PReal	Reale	5.5	5.5
PTime	Ora	23:05:01	t{'23:05:01'}
PDate	Data	2011-03-02	d{'2011-03-02'}
PTimeStamp	Timestamp	2011-03-02 23:05:01	ts{'2011-03-02 23:05:01'}
PColumn	Sconosciuto	IntValue	IntValue

Nella query SQL, i parametri del flusso vengono specificati come in un'espressione CLEM, mediante '\$P-<parameter_name>', dove <parameter_name> è il nome definito per il parametro del flusso.

Quando si fa riferimento a un campo, è necessario definire il tipo di archiviazione come Sconosciuto, mentre il valore del parametro deve essere racchiuso tra virgolette, se necessario. Di conseguenza, in base agli esempi riportati nella tabella, se è stata immessa la query SQL:

```
select "IntValue" from Table1 where "IntValue" < '$P-PInt';
```

la query verrà valutata come:

```
select "IntValue" from Table1 where "IntValue" < 5;
```

Se è necessario fare riferimento al campo IntValue mediante il parametro PColumn, per ottenere lo stesso risultato è necessario specificare la query nel modo riportato di seguito:

```
select "IntValue" from Table1 where "'$P-PColumn'" < '$P-PInt';
```

Nodo Testo variabile

È possibile utilizzare i nodi Testo variabile per leggere dati da file di testo a campi liberi, ovvero file i cui record contengono un numero costante di campi e un numero variabile di caratteri, noti anche come file di testo delimitati. Questo tipo di nodi può essere utilizzato per file con testo di intestazione a lunghezza fissa e alcuni tipi di annotazioni. I record vengono letti uno alla volta e passati nel flusso finché non viene letto l'intero file.

Note per la lettura dei dati di testo delimitati

- Alla fine di ogni riga, i record devono essere delimitati da un carattere di nuova riga. Il carattere di nuova riga non deve essere utilizzato per altri scopi (per esempio, all'interno di un nome di campo o di un valore). Gli spazi iniziali e finali andrebbero eliminati per risparmiare spazio, anche se questo non è fondamentale. Se necessario, tali spazi possono essere eliminati dal nodo.
- I campi devono essere delimitati da una virgola o da un altro carattere che andrebbe utilizzato solo come delimitatore, cioè non deve comparire all'interno di nomi di campo o di valori. Se ciò non è possibile, tutti i campi testo possono essere racchiusi tra virgolette, purché nessuno dei nomi dei campi o dei valori di testo contenga virgolette. Se i nomi dei campi o i valori contengono virgolette, in alternativa i campi testo possono essere racchiusi tra virgolette singole, sempre a condizione che le virgolette singole non siano utilizzate in altri punti all'interno di valori. Se non è possibile utilizzare né virgolette singole, né virgolette, i valori di testo devono essere modificati in modo da rimuovere o sostituire il carattere di delimitazione oppure le virgolette o le virgolette singole.
- Ogni riga, intestazione compresa, deve contenere lo stesso numero di campi.
- La prima riga deve contenere i nomi dei campi. In caso contrario, deselezionare l'opzione **Leggi i nomi dei campi dal file** per assegnare a ogni campo un nome generico come *Campo1*, *Campo2* e così via.
- La seconda riga deve contenere il primo record di dati. Non devono essere presenti righe vuote o commenti.
- I valori numerici non devono includere il separatore delle migliaia oppure il simbolo di raggruppamento — ad esempio, in 3,000.00 non deve essere presente la virgola. L'indicatore dei decimali (virgola o punti negli Stati Uniti e in Gran Bretagna) deve essere usato solo ove appropriato.
- I valori di data e ora devono essere in uno dei formati riconosciuti nella finestra di dialogo Opzioni flusso, per esempio GG/MM/AAAA o HH:MM:SS. Tutti i campi data e ora nel file dovrebbero seguire lo stesso formato e i campi contenenti una data devono utilizzare lo stesso formato per tutti i valori all'interno di quel campo.

Impostazione delle opzioni relative al nodo Testo variabile

Le opzioni vengono impostate nella scheda File della finestra di dialogo del nodo Testo variabile.

File. Specificare il nome del file. È possibile immettere un nome di file o fare clic sul pulsante con i puntini di sospensione (...) per selezionare un file. Il percorso del file viene mostrato quando il file viene selezionato e i relativi contenuti separati da delimitatori vengono visualizzati nel riquadro sottostante.

Il testo di esempio visualizzato dalla sorgente dati può essere copiato e incollato nei controlli seguenti: caratteri di commento EOL e delimitatori definiti dall'utente. Utilizzare Ctrl-C e Ctrl-V per copiare e incollare.

Leggi i nomi dei campi dal file. Selezionata per default, questa opzione considera la prima riga nel file di dati come etichette per la colonna. Se la prima riga non è una riga di intestazione, deselezionare questa opzione per assegnare automaticamente a ogni campo un nome generico, per esempio *Campo1*, *Campo2*, in base al numero di campi nell'insieme di dati.

Specifica numero di campi. Specificare il numero di campi in ogni record. Il numero di campi può essere rilevato automaticamente a condizione che i record siano delimitati da un carattere di nuova riga. È anche possibile impostare manualmente un numero.

Ignora caratteri intestazione. Specificare il numero di caratteri che si desidera ignorare all'inizio del primo record.

Caratteri di commento EOL. Specificare caratteri quali # o ! per indicare annotazioni nei dati. Ogni volta che nel file di dati verrà rilevato uno di questi caratteri, tutti gli elementi precedenti il carattere di nuova riga (escluso tale carattere) verranno ignorati.

Elimina spazi all'inizio e alla fine. Selezionare le opzioni per l'eliminazione degli spazi iniziali e finali nelle stringhe durante l'importazione.

Nota: i confronti tra stringhe che utilizzano e non utilizzano il push back SQL potrebbero generare risultati diversi in presenza di spazi finali.

Caratteri non validi. Selezionare **Scarta** per rimuovere i caratteri non validi dall'origine dei dati. Selezionare **Sostituisci con** per sostituire i caratteri non validi con il simbolo specificato (un solo carattere). I caratteri non validi sono caratteri null o caratteri non esistenti nel metodo di codifica specificato.

Codifica. Specifica il metodo di codifica del testo utilizzato. È possibile scegliere tra l'impostazione di default del sistema, l'impostazione di default del flusso o UTF-8.

- L'impostazione di default del sistema è specificata nel Pannello di controllo di Windows oppure, se si esegue il programma in modalità distribuita, sul computer server.
- L'impostazione di default del flusso è specificata nella finestra di dialogo Proprietà flusso.

Simbolo decimale. Selezionare il tipo di separatore decimale utilizzato nell'origine dati. **Default flusso** rappresenta il carattere selezionato nella scheda Opzioni della finestra di dialogo delle proprietà del flusso. In alternativa, selezionare **Punto (.)** o **Virgola (,)** per utilizzare il relativo simbolo come separatore di decimali per tutti i dati inclusi nella finestra di dialogo.

Il delimitatore di riga è il carattere di nuova riga. Per utilizzare il carattere di nuova riga come delimitatore di riga, invece di un delimitatore di campo, selezionare questa opzione. Ad esempio, questa opzione può essere utile se in una riga è presente un numero dispari di delimitatori che forza un a capo. La selezione di questa opzione è incompatibile con la selezione di **Nuova riga** nell'elenco Delimitatori.

Considerare che se si seleziona questa opzione, qualsiasi valore vuoto alla fine della riga dei dati verrà eliminato.

Delimitatori. Tramite le caselle di controllo incluse in questo controllo è possibile specificare i caratteri da utilizzare come delimitatori per i campi del file, per esempio la virgola (,). È possibile specificare più delimitatori, per esempio ", |", per i record che utilizzano più delimitatori. Il delimitatore di default è la virgola.

Nota: se la virgola è definita anche come separatore decimale, le impostazioni di default definite in questo punto non hanno effetto. Nei casi in cui la virgola è sia delimitatore di campo che separatore decimale, selezionare **Altro** nell'elenco Delimitatori, quindi specificare manualmente una virgola nel campo di immissione.

Selezionare **Consenti più delimitatori vuoti** per trattare più caratteri di delimitazione vuoti adiacenti come un unico delimitatore. Per esempio, se un valore nei dati è seguito da quattro spazi e quindi da un altro valore, verranno considerati due campi anziché cinque.

Righe da esaminare per colonna e tipo. Specificare il numero di righe e colonne in cui ricercare i tipi di dati specificati.

Riconosci automaticamente date e ore. Per abilitare IBM SPSS Modeler in modo da provare a riconoscere automaticamente le voci di dati come date e ore, selezionare questa casella di controllo. Per esempio, questo significa che una voce come 07-11-1965 verrà identificata come una data e 02:35:58 come ora. Tuttavia, le voci ambigue, come 07111965 o 023558, verranno identificate come numeri interi in quanto non ci sono delimitatori fra i numeri.

Nota: per evitare potenziali problemi relativi ai dati durante l'utilizzo dei file di dati da versioni precedenti di IBM SPSS Modeler, questa casella è disattivata per default per le informazioni salvate nelle versioni precedenti alla 13.

Virgolette. Tramite gli elenchi a discesa è possibile specificare in che modo si desidera vengano utilizzate le virgolette singole e doppie durante l'importazione. È possibile selezionare **Scarta** per scartare tutti i tipi di virgolette, **Includi come testo** per includerle nei valori di campo o **Accoppia e scarta** per individuare coppie di virgolette corrispondenti e quindi scartarle. Se per una virgoletta non viene trovato il corrispondente, verrà visualizzato un messaggio di errore. Le opzioni **Scarta** e **Accoppia e scarta** archiviano il valore di campo (senza virgolette) come stringa.

Mentre si utilizza questa finestra di dialogo è sempre possibile fare clic su **Aggiorna** per ricaricare i campi dalla sorgente dati. Questa opzione risulta utile quando si modificano le connessioni dei dati al nodo origine o quando si passa da una scheda all'altra nella finestra di dialogo.

Nodo Testo fisso

È possibile utilizzare i nodi Testo fisso per importare dati da file di testo a campi fissi (ovvero file i cui campi non sono delimitati, iniziano nella stessa posizione e sono di lunghezza fissa). Nel formato a campi fissi vengono in genere archiviati dati di versioni precedenti o generati dalla macchina. Tramite la scheda File del nodo Testo fisso è possibile specificare la posizione e la lunghezza delle colonne nei dati.

Impostazione delle opzioni relative al nodo Testo fisso

La scheda File del nodo Testo fisso consente di importare dati in IBM SPSS Modeler e di specificare la posizione delle colonne e la lunghezza dei record. È possibile fare clic all'interno del riquadro di anteprima dei dati al centro della finestra di dialogo per aggiungere frecce indicanti i punti di separazione tra i campi.

File. Specificare il nome del file. È possibile immettere un nome di file o fare clic sul pulsante con i puntini di sospensione (...) per selezionare un file. Dopo avere selezionato un file, viene visualizzato il relativo percorso e il contenuto con i delimitatori viene visualizzato nel riquadro sottostante.

Il riquadro di anteprima dei dati può essere utilizzato per specificare la lunghezza e la posizione delle colonne. Il righello nella parte superiore del riquadro di anteprima consente di misurare la lunghezza delle variabili e di specificare il punto di separazione tra le variabili. È possibile specificare le linee relative ai punti di separazione facendo clic nell'area del righello al di sopra dei campi. I punti di separazione possono essere spostati tramite trascinamento e scartati trascinandoli all'esterno dell'area di anteprima dei dati.

- Per ogni punto di separazione viene aggiunto automaticamente un nuovo campo ai campi della tabella sottostante.
- Le posizioni iniziali indicate dalle frecce vengono aggiunte automaticamente alla colonna Inizio nella tabella sottostante.

Ignora car. nuova riga. Selezionare se si desidera ignorare il carattere di nuova riga alla fine di ciascun record.

Ignora righe intestazione. Specificare il numero di righe che si desidera ignorare all'inizio del primo record. Questa opzione può essere utilizzata per ignorare le intestazioni delle colonne.

Lunghezza record. Specificare il numero di caratteri in ogni record.

Campo. Sono elencati tutti i campi definiti per questo file di dati. Esistono due modi per definire i campi:

- Definizione interattiva dei campi mediante il riquadro di anteprima dei dati.
- Definizione manuale dei campi mediante l'aggiunta di righe di campo vuote nella tabella. Fare clic sul pulsante a destra del riquadro dei campi per aggiungere nuovi campi. Quindi immettere nel campo vuoto un nome di campo, una posizione iniziale e una lunghezza. L'impostazione di queste opzioni determina l'aggiunta automatica delle frecce al riquadro di anteprima, con la possibilità di modifiche successive.

Per rimuovere un campo definito in precedenza, selezionarlo nell'elenco e fare clic sul pulsante con la X rossa.

Inizio. Specificare la posizione del primo carattere nel campo. Per esempio, se il secondo campo di un record inizia in corrispondenza del sedicesimo carattere, come posizione iniziale si specificherà 16.

Lunghezza. Specificare il numero di caratteri presenti nel valore più lungo per ciascun campo. In questo modo viene definito il punto di interruzione per il campo successivo.

Elimina spazi all'inizio e alla fine. Selezionare per eliminare gli spazi iniziali e finali nelle stringhe durante l'importazione.

Nota: i confronti tra stringhe che utilizzano e non utilizzano il push back SQL potrebbero generare risultati diversi in presenza di spazi finali.

Caratteri non validi. Selezionare **Scarta** per rimuovere i caratteri non validi dall'input di dati. Selezionare **Sostituisci con** per sostituire i caratteri non validi con il simbolo specificato (un solo carattere). I caratteri non validi sono caratteri null (0) o caratteri non esistenti nella codifica corrente.

Codifica. Specifica il metodo di codifica del testo utilizzato. È possibile scegliere tra l'impostazione di default del sistema, l'impostazione di default del flusso o UTF-8.

- L'impostazione di default del sistema è specificata nel Pannello di controllo di Windows oppure, se si esegue il programma in modalità distribuita, sul computer server.
- L'impostazione di default del flusso è specificata nella finestra di dialogo Proprietà flusso.

Simbolo decimale. Selezionare il tipo di separatore decimale utilizzato nell'origine dati. **Default flusso** rappresenta il carattere selezionato nella scheda Opzioni della finestra di dialogo delle proprietà del flusso. In alternativa, selezionare **Punto (.)** o **Virgola (,)** per utilizzare il relativo simbolo come separatore di decimali per tutti i dati inclusi nella finestra di dialogo.

Riconosci automaticamente date e ore. Per abilitare IBM SPSS Modeler in modo da provare a riconoscere automaticamente le voci di dati come date e ore, selezionare questa casella di controllo. Per esempio, questo significa che una voce come 07-11-1965 verrà identificata come una data e 02:35:58 come ora. Tuttavia, le voci ambigue, come 07111965 o 023558, verranno identificate come numeri interi in quanto non ci sono delimitatori fra i numeri.

Nota: per evitare potenziali problemi relativi ai dati durante l'utilizzo dei file di dati da versioni precedenti di IBM SPSS Modeler, questa casella è disattivata per default per le informazioni salvate nelle versioni precedenti alla 13.

Righe da esaminare per il tipo. Indicare il numero di righe da esaminare per i tipi di dati specificati.

Mentre si utilizza questa finestra di dialogo è sempre possibile fare clic su **Aggiorna** per ricaricare i campi dalla sorgente dati. Questa opzione risulta utile quando si modificano le connessioni dei dati al nodo origine o quando si passa da una scheda all'altra nella finestra di dialogo.

Impostazione dell'archiviazione e della formattazione dei campi

Le opzioni nella scheda Dati, comuni ai nodi Testo fisso, Testo variabile, Origine XML e Input utente, consentono di specificare il tipo di archiviazione dei campi importati o creati in IBM SPSS Modeler. Per i nodi Testo fisso, Testo variabile e Input utente è possibile specificare anche la formattazione dei campi e altri metadati.

Per i dati letti da altre sorgenti, l'archiviazione viene determinata automaticamente ma può essere modificata utilizzando una funzione di conversione, quale `to_integer`, in un nodo Riempimento o Ricava.

Campo. Utilizzare la colonna *Campo* per visualizzare e selezionare i campi nell'insieme di dati corrente.

Sovrascrivi. Selezionare la casella di controllo nella colonna *Sovrascrivi* per attivare le opzioni nelle colonne *Archiviazione* e *Formato di input*.

Archiviazione dei dati

L'archiviazione definisce le modalità di archiviazione dei dati in un campo. Per esempio, un campo con valori 1 e 0 è un campo in cui sono archiviati dati di tipo intero. Si distingue dal livello di misurazione, che descrive l'utilizzo dei dati e non influisce sull'archiviazione. Per esempio, è possibile impostare il livello di misurazione per un campo intero con valori 1 e 0 su *Flag*. In genere, questo indica che 1=*Vero* e 0=*Falso*. Mentre l'archiviazione deve essere stabilita all'origine, il livello di misurazione può essere modificato tramite un nodo Tipo in qualsiasi punto del flusso. Per ulteriori informazioni, consultare l'argomento "Livelli di misurazione" a pagina 112.

I tipi di archiviazione disponibili sono:

- **Stringa.** Utilizzato per i campi che contengono dati non numerici, detti anche dati alfanumerici. Una stringa può includere qualsiasi sequenza di caratteri, quale *mario*, *Classe 2* o *1234*. Si noti che i numeri presenti nelle stringhe non possono essere utilizzati nei calcoli.
- **Intero.** Un campo i cui valori sono numeri interi.
- **Reale.** I valori sono numeri che possono includere decimali (non limitati ai numeri interi). Il formato di visualizzazione viene specificato nella finestra di dialogo Proprietà flusso e può essere ignorato per singoli campi in un nodo Tipo (scheda Formato).
- **Data.** Valori di data specificati in un formato standard quale anno, mese e giorno (per esempio, 26.09.07). Il formato specifico è indicato nella finestra di dialogo Proprietà flusso.
- **Ora.** Tempo misurato come durata. Per esempio, se una chiamata di assistenza è durata 1 ora, 26 minuti e 38 secondi, potrà essere rappresentata come 01:26:38, in base al formato dell'ora corrente specificato nella finestra di dialogo Proprietà stream.
- **Timestamp.** Valori che includono sia la data che l'ora, ad esempio 2007-09-26 09:04:00, che dipendono dai formati di data ed ora correnti specificati nella finestra di dialogo Proprietà flusso. Notare che potrebbe essere necessario racchiudere i valori timestamp tra doppi apici per garantire che siano interpretati come valore singolo piuttosto che come valori di data ed ora separati. Ad esempio quando si immettono valori in un nodo Input utente.

Conversioni della classe di archiviazione. È possibile convertire la classe di archiviazione di un campo utilizzando una vasta gamma di funzioni di conversione, per esempio `to_string` e `to_integer`, in un nodo Riempimento. Per ulteriori informazioni, consultare l'argomento "Conversione della classe di archiviazione mediante il nodo Riempimento" a pagina 133. Notare che le funzioni di conversione (e tutte le altre funzioni che richiedono un tipo specifico di input, come un valore data o ora) dipendono dai formati correnti specificati nella finestra di dialogo Proprietà flusso. Per esempio, se si desidera convertire un campo stringa con valori *Gen 2003*, *Feb 2003* e così via in campo di archiviazione della data, selezionare **MES AAAA** come formato di data di default per il flusso. Le funzioni di conversione sono disponibili anche dal nodo Ricava per la conversione temporanea durante un calcolo di derivazione. È

inoltre possibile utilizzare il nodo Ricava per eseguire altre manipolazioni, quali ricodifiche di campi stringa con valori categoriali. Per ulteriori informazioni, consultare l'argomento "Ricodifica di valori con il nodo Ricava" a pagina 132.

Lettura di dati misti. Quando si leggono campi con archiviazione numerica (numeri interi, reali, ora, timestamp o data), gli eventuali valori non numerici vengono impostati come valori null o valori mancanti di sistema poiché, a differenza di altre applicazioni, IBM SPSS Modeler non consente di utilizzare tipi di archiviazione misti all'interno dello stesso campo. Per evitare questo inconveniente, è necessario modificare opportunamente il tipo di archiviazione nel nodo origine o nell'applicazione esterna per fare in modo che tutti i campi contenenti dati misti siano letti come stringhe.

Formato di input del campo (solo nodi Testo fisso, Testo variabile e Input utente)

Per tutti i tipi di classe di archiviazione eccetto Stringa e Intero, è possibile specificare le opzioni di formattazione per il campo selezionato utilizzando l'elenco a discesa. Per esempio, quando si uniscono dati tra diverse opzioni internazionali, potrebbe essere necessario specificare il punto (.) come separatore decimale per un campo e la virgola per un altro campo.

Le opzioni di input definite nel nodo origine sovrascrivono le opzioni di formattazione specificate nella finestra di dialogo delle proprietà del flusso, ma non vengono mantenute successivamente nel flusso. Queste opzioni sono state progettate per analizzare correttamente l'input in base alle informazioni note sui dati. I formati specificati risultano utili per analizzare i dati mentre vengono letti in IBM SPSS Modeler e non per stabilire in quale modo devono essere formattati dopo essere stati letti in IBM SPSS Modeler. Per definire il formato di ognuno dei campi in un'altra posizione del flusso, utilizzare la scheda Formato di un nodo Tipo. Per ulteriori informazioni, consultare l'argomento "Scheda per l'impostazione dei formati dei campi" a pagina 120.

Le opzioni variano in base al tipo di archiviazione. Per esempio, per il tipo Reale, è possibile selezionare **Punto (.)** o **Virgola (,)** come separatore decimale. Per i campi Timestamp, quando si seleziona **Specifica** dall'elenco a discesa viene aperta un'altra finestra di dialogo. Per ulteriori informazioni, consultare l'argomento "Impostazione delle opzioni relative al formato dei campi" a pagina 121.

Per tutti i tipi di archiviazione, è inoltre possibile selezionare **Default flusso** per utilizzare le impostazioni predefinite del flusso per l'importazione. Le impostazioni del flusso sono definite nella finestra di dialogo delle proprietà del flusso.

Opzioni aggiuntive

Nella scheda Dati è possibile definire molte altre opzioni:

- Per visualizzare le impostazioni di archiviazione per dati non più connessi al nodo corrente (per esempio dati di addestramento), selezionare **Visualizza impostazioni campi inutilizzati**. È possibile eliminare i campi precedenti facendo clic su **Cancella**.
- Mentre si utilizza questa finestra di dialogo è sempre possibile fare clic su **Aggiorna** per ricaricare i campi dalla sorgente dati. Questa opzione risulta utile quando si modificano le connessioni dei dati al nodo origine o quando si passa da una scheda all'altra nella finestra di dialogo.

Nodo Data Collection

I nodi di origine di Data Collection importano i dati di indagine in base a IBM SPSS Data Collection Survey Reporter Developer Kit utilizzato dal software di ricerca di mercato di IBM Corp.. Tale formato distingue i **dati del caso** — le risposte effettive alle domande raccolte durante un'indagine — dai **metadati** che descrivono il modo in cui i dati del caso vengono raccolti ed organizzati. I metadati sono costituiti da informazioni quali testi di domande, nomi di variabili e descrizioni, definizioni di variabili a risposta multipla, traduzioni delle stringhe di testo, nonché dalla definizione della struttura dei dati del caso.

Nota: questo nodo richiede IBM SPSS Data Collection Survey Reporter Developer Kit, distribuito insieme ai prodotti software IBM SPSS Data Collection da IBM Corp.. Per ulteriori informazioni consultare la pagina Web IBM SPSS Data Collection all'indirizzo . Fatta eccezione per l'installazione del Developer Kit, non è necessario eseguire altre operazioni di configurazione.

Commenti

- I dati di indagine vengono letti dal formato piatto tabulare VDATA o da sorgenti nel formato gerarchico HDATA se includono una sorgente di metadati (è necessario IBM SPSS Data Collection versione 4.5 o superiore).
- I tipi vengono istanziati automaticamente utilizzando le informazioni fornite dai metadati.
- Quando si importano in IBM SPSS Modeler dati di indagine, le domande vengono interpretate come campi, con un record per ogni rispondente.

Opzioni scheda File del nodo Importazione di Data Collection

La scheda File del nodo Data Collection consente di specificare le opzioni relative ai metadati e ai dati del caso che si desidera importare.

Impostazioni dei metadati

Nota: per visualizzare l'elenco completo dei tipi di file di provider disponibili, è necessario installare IBM SPSS Data Collection Survey Reporter Developer Kit, disponibile con il software IBM SPSS Data Collection. Per ulteriori informazioni, vedere la pagina Web di IBM SPSS Data Collection all'indirizzo <http://www.ibm.com/software/analytics/spss/products/data-collection/survey-reporter-dev-kit/>

Provider di metadati. I dati dell'indagine possono essere importati da diversi formati supportati dal software IBM SPSS Data Collection Survey Reporter Developer Kit. I tipi di provider disponibili includono:

- **DataCollectionMDD.** Legge i metadati da un file di definizione del questionario (*.mdd*). Questo è il formato standard del Modello dati di IBM SPSS Data Collection.
- **Database ADO.** Legge i dati e i metadati del caso dai file ADO. Specificare il nome e la posizione del file *.adoinfo* che contiene i metadati. Il nome interno di questo DSC è *mrADODsc*.
- **Database In2data.** Legge i dati e i metadati del caso In2data. Il nome interno di questo DSC è *mrI2dDsc*.
- **File di registro Data Collection.** Legge i metadati da un file di log IBM SPSS Data Collection standard. In genere, i file di registro presentano un'estensione *.tmp*. Tuttavia, alcuni file di registro possono avere un'estensione diversa. Se necessario, è possibile rinominare il file in modo che presenti un'estensione *.tmp*. Il nome interno di questo DSC è *mrLogDsc*.
- **File di definizione Quancept.** Converte i metadati in script Quancept. Specificare il nome del file *.qdi* Quancept. Il nome interno di questo DSC è *mrQdiDrsDsc*.
- **Database Quanvert.** Legge i dati e i metadati del caso Quanvert. Specificare il nome e la posizione del file *.qvinfo* o *.pkd*. Il nome interno di questo DSC è *mrQvDsc*.
- **Data Collection Participation Database.** Legge le tabelle Campione e Tabella cronologia di un progetto e crea le variabili categoriali derivate corrispondenti alle colonne di queste tabelle. Il nome interno di questo DSC è *mrSampleReportingMDSC*.
- **File Statistics.** Legge i metadati ed i dati del caso da un file *.sav* di IBM SPSS Statistics. Scrive i dati del caso in un file IBM SPSS Statistics *.sav* per l'analisi in IBM SPSS Statistics. Scrive i metadati da un file *.sav* di IBM SPSS Statistics ad un file *.mdd*. Il nome interno di questo DSC è *mrSavDsc*.
- **File Surveycraft.** Legge i dati e i metadati del caso SurveyCraft. Specificare il nome del file *.vq* SurveyCraft. Il nome interno di questo DSC è *mrSCDsc*.
- **File di script Data Collection.** Legge dai metadati in un file *mrScriptMetadata*. In genere, questi file presentano un'estensione *.mdd* o *.dms*. Il nome interno di questo DSC è *mrScriptMDSC*.

- **File XML Triple-S.** Legge i metadati da un file Triple-S in formato XML. Il nome interno di questo DSC è *mrTripleSDsc*.

Proprietà dei metadati. Se lo si desidera, selezionare **Proprietà** per specificare la versione dell'indagine da importare e la lingua, il contesto e il tipo di etichetta da utilizzare. Per ulteriori informazioni, consultare l'argomento "Proprietà dei metadati dell'importazione di IBM SPSS Data Collection" a pagina 28.

Impostazione dei dati del caso

Nota: per visualizzare l'elenco completo dei tipi di file di provider disponibili, è necessario installare IBM SPSS Data Collection Survey Reporter Developer Kit, disponibile con il software IBM SPSS Data Collection. Per ulteriori informazioni, vedere la pagina Web di IBM SPSS Data Collection all'indirizzo <http://www.ibm.com/software/analytics/spss/products/data-collection/survey-reporter-dev-kit/>

Ottieni impostazione dei dati del caso. Solo in caso di lettura di metadati da file *.mdd*, fare clic su **Ottieni impostazione dei dati del caso** per determinare quali sorgenti dei dati del caso sono associate ai metadati selezionati, oltre alle impostazioni specifiche necessarie per accedere a una determinata sorgente. Questa opzione è disponibile solo per i file *.mdd*.

Provider di dati del caso. Sono supportati i seguenti tipi di provider:

- **Database ADO.** Legge i dati del caso utilizzando l'interfaccia Microsoft ADO. Selezionare UDL OLE-DB per il tipo di dati del caso e specificare una stringa di connessione nel campo UDL di dati del caso. Per ulteriori informazioni, consultare l'argomento "Stringa di connessione al database" a pagina 29. Il nome interno di questo componente è *mrADODsc*.
- **File di testo delimitato (Excel).** Legge i dati del caso da un file delimitato da virgola (.CSV), quale un file generato da Excel. Il nome interno è *mrCsvDsc*.
- **File di dati Data Collection.** Legge i dati del caso da un file in formato dati di IBM SPSS Data Collection nativo (a partire da IBM SPSS Data Collection versione 4.5). Il nome interno è *mrDataFileDsc*.
- **Database In2data.** Legge i dati e i metadati del caso da un file di database (*.i2d*) In2data. Il nome interno è *mrI2dDsc*.
- **File di registro Data Collection.** Legge i dati del caso da un file di registro IBM SPSS Data Collection standard. In genere, i file di registro presentano un'estensione *.tmp*. Tuttavia, alcuni file di registro possono avere un'estensione diversa. Se necessario, è possibile rinominare il file in modo che presenti un'estensione *.tmp*. Il nome interno è *mrLogDsc*.
- **File di dati Quantum.** Legge i dati del caso da un file ASCII (*.dat*) formato Quantum. Il nome interno è *mrPunchDsc*.
- **File di dati Quancept.** Legge i dati del caso da un file *.drs*, *.drz* o *.dru* Quancept. Il nome interno è *mrQdiDrsDsc*.
- **Database Quanvert.** Legge i dati del caso da un file *.pkd* o *qvinfo* Quanvert. Il nome interno è *mrQvDsc*.
- **Database Data Collection (MS SQL Server).** Legge i dati del caso da un database Microsoft SQL Server relazionale. Per ulteriori informazioni, consultare l'argomento "Stringa di connessione al database" a pagina 29. Il nome interno è *mrRdbDsc2*.
- **File Statistics.** Legge i dati del caso da un file *.sav* di IBM SPSS Statistics. Il nome interno è *mrSavDsc*.
- **File Surveycraft.** Legge i dati del caso da un file *.qdt* SurveyCraft. I file *.vq* e *.qdt* devono trovarsi nella stessa directory e consentire entrambi l'accesso in lettura e in scrittura. Questo non è il modo di default in cui vengono creati quando si usa SurveyCraft, pertanto per importare i dati SurveyCraft è necessario spostare uno dei file. Il nome interno è *mrScDsc*.
- **File di dati Triple-S.** Legge i dati del caso da un file di dati Triple-S, sia nel formato a lunghezza fissa che nel formato delimitato da virgola. Il nome interno è *mr TripleDsc*.
- **Data Collection XML.** Legge i dati del caso da un file di dati XML IBM SPSS Data Collection. In genere, questo formato può essere utilizzato per trasferire i dati del caso da una posizione a un'altra. Il nome interno è *mrXmlDsc*.

Tipo di dati del caso. Specifica se il caso viene letto da un file, cartella, UDL OLE-DB o DSN ODBC e aggiorna di conseguenza le opzioni della finestra di dialogo. Le opzioni valide dipendono dal tipo di provider. Per i provider di database, è possibile specificare le opzioni per la connessione OLE-DB o ODBC. Per ulteriori informazioni, consultare l'argomento "Stringa di connessione al database" a pagina 29.

Progetto di dati del caso. Durante la lettura dei dati del caso da un database IBM SPSS Data Collection, è possibile immettere il nome del progetto. Per tutti gli altri tipi di dati del caso, questa impostazione deve essere lasciata vuota.

Importazione variabili

Importa variabili di sistema. Specifica se vengono importate le variabili di sistema, incluse le variabili che indicano lo stato dell'intervista (in corso, completata, data fine e così via). È possibile scegliere **Nessuna**, **Tutte** o **Comuni**.

Importa variabili "Codes". Controlla l'importazione delle variabili che rappresentano i codici utilizzati per le risposte aperte "Altro" per le variabili categoriali.

Importa variabili "SourceFile". Controlla l'importazione di variabili che contengono nomi file o immagini delle risposte analizzate.

Importa variabili a risposta multipla come. Le variabili a risposta multipla possono essere importate come campi flag multipli (un insieme dicotomico multiplo), che è il metodo di default per i nuovi stream. I flussi creati nelle versioni di IBM SPSS Modeler precedenti alla 12.0 importavano le risposte multiple in un unico campo, con i valori separati da virgole. Il metodo precedente è ancora supportato per consentire l'esecuzione dei flussi esistenti come in precedenza, ma si consiglia di aggiornare i vecchi stream per l'utilizzo del nuovo metodo. Per ulteriori informazioni, consultare l'argomento "Importazione di insiemi di risposte multiple" a pagina 29.

Proprietà dei metadati dell'importazione di IBM SPSS Data Collection

Durante l'importazione dei dati di indagine IBM SPSS Data Collection, nella finestra di dialogo Proprietà dei metadati, è possibile specificare la versione dell'indagine da importare e la lingua, il contesto ed il tipo di etichetta da utilizzare. Si noti che è possibile importare solo una lingua, un contesto e un tipo di etichetta alla volta.

Versione. Ogni versione dell'indagine può essere considerata come una snapshot dei metadati utilizzati per raccogliere un particolare insieme di dati del caso. Via via che un questionario viene modificato, possono essere create più versioni. È possibile importare la versione più recente, tutte le versioni o una versione specifica.

- **Tutte le versioni.** Selezionare questa opzione se si desidera utilizzare una combinazione (superinsieme) di tutte le versioni disponibili (denominata, talvolta, superversione). In caso di conflitti tra le versioni, le più recenti prevalgono generalmente sulle meno recenti. Per esempio, se un'etichetta categoria è diversa in una delle versioni, verrà utilizzato il testo dell'ultima versione.
- **Ultima versione.** Selezionare questa opzione se si desidera utilizzare la versione più recente.
- **Specificata versione.** Selezionare questa opzione se si desidera utilizzare una versione specifica dell'indagine.

L'utilizzo di tutte le versioni è utile quando, per esempio, si desidera esportare i dati del caso per più di una versione e sono state apportate modifiche alle definizioni di categorie e variabili, cioè i dati del caso raccolti con una versione non sono validi in un'altra. Selezionando tutte le versioni per le quali si desidera esportare i dati del caso, è possibile generalmente esportare i dati del caso raccolti con le diverse versioni contemporaneamente senza incorrere in errori di validità dovuti alle differenze tra le versioni. Tuttavia, a seconda delle modifiche nelle versioni, è possibile che si verifichino ancora errori di validità.

Lingua. Le domande e il testo correlato possono essere archiviati in più lingue nei metadati. È possibile utilizzare la lingua di default dell'indagine o specificare una lingua particolare. Se un elemento non è disponibile nella lingua specificata, verrà utilizzata quella di default.

Contesto. Selezionare il contesto dell'utente che si desidera utilizzare. Il contesto dell'utente determina quali testi vengono visualizzati. Per esempio, selezionare **Domanda** per visualizzare i testi delle domande o **Analisi** per visualizzare testi più brevi adatti alla visualizzazione durante l'analisi dei dati.

Tipo di etichetta. Elenca i tipi di etichette che sono stati definiti. L'impostazione di default è **etichetta**, che viene utilizzata per i testi delle domande nel contesto dell'utente Domanda e le descrizioni variabili nel contesto dell'utente Analisi. È possibile definire altri tipi di etichette per istruzioni, descrizioni e così via.

Stringa di connessione al database

Quando viene utilizzato il nodo IBM SPSS Data Collection per importare i dati del caso da un database mediante OLE-DB o ODBC, selezionare **Modifica** dalla scheda File per accedere alla finestra di dialogo Stringa di connessione, che consente di personalizzare la stringa di connessione passata la provider per ottimizzare la connessione.

Proprietà avanzate

Quando si utilizza il nodo IBM SPSS Data Collection per importare i dati del caso da un database che richiede un login esplicito, selezionare **Avanzate** per fornire un ID utente e una password per accedere alla sorgente dati.

Importazione di insiemi di risposte multiple

Le variabili a risposta multipla possono essere importate da IBM SPSS Data Collection come insiemi dicotomici multipli, con un campo flag separato per ogni valore possibile della variabile. Per esempio, se agli intervistati viene chiesto di selezionare da un elenco i musei che hanno visitato, l'insieme includerà un campo flag separato per ogni museo elencato.

Dopo l'importazione dei dati, è possibile aggiungere o modificare gli insiemi di risposte multiple di qualsiasi nodo che include una scheda Filtro. Per ulteriori informazioni, consultare l'argomento "Modifica di insiemi di risposte multiple" a pagina 124.

Importazione di risposte multiple in un unico campo (per i flussi creati nelle versioni precedenti)

Nelle precedenti versioni di IBM SPSS Modeler, l'importazione delle risposte multiple non si effettuava come descritto sopra ma in un unico campo, con i valori separati da virgole. Questo metodo è tuttora supportato per i flussi esistenti, ma si consiglia di aggiornare tali stream per l'utilizzo del nuovo metodo.

Note sull'importazione di colonne in IBM SPSS Data Collection

Le colonne di dati di IBM SPSS Data Collection vengono lette in IBM SPSS Modeler come riepilogato nella tabella seguente.

Tabella 2. Riepilogo importazione colonna IBM SPSS Data Collection

Tipo di colonna IBM SPSS Data Collection	Archiviazione IBM SPSS Modeler	Livello di misurazione
Flag booleano (sì/no)	Stringa	Flag (valori 0 e 1)
Categoriale	Stringa	Nominale
Data o timestamp	Timestamp	Continuo
Doppio (valore a virgola mobile entro un intervallo specificato)	Reale	Continuo

Tabella 2. Riepilogo importazione colonna IBM SPSS Data Collection (Continua)

Tipo di colonna IBM SPSS Data Collection	Archiviazione IBM SPSS Modeler	Livello di misurazione
Lungo (valore intero entro un intervallo specificato)	Intero	Continuo
Testo (descrizione testo libero)	Stringa	Typeless
Livello (indica griglie o cicli all'interno di una domanda)	Non si verifica in VDATA e non è importato in IBM SPSS Modeler	
Oggetto (dati binari quali fax con testo scritto a mano o registrazioni vocali)	Non importato in IBM SPSS Modeler	
Nessuno (tipo sconosciuto)	Non importato in IBM SPSS Modeler	
Colonna Respondent.Serial (associa un ID univoco a ogni rispondente)	Intero	Typeless

Per evitare possibili incoerenze tra le etichette di valori lette dai metadati e i valori effettivi, tutti i valori dei metadati vengono convertiti in caratteri minuscoli. Ad esempio, l'etichetta di valore *E1720_years* viene convertita in *e1720_years*.

Analytic Server Origine

L'origine Analytic Server consente di eseguire un flusso su HDFS (Hadoop Distributed File System). Le informazioni contenute in un'origine dati Analytic Server possono provenire da una varietà di fonti, incluso:

- File di testo su HDFS
- Database
- HCatalog

In genere un flusso con un'origine Analytic Server verrà eseguito su HDFS; tuttavia, se un flusso contiene un nodo che non è supportato per l'esecuzione su HDFS, la maggiore quantità di flusso possibile verrà "rimandata" a Analytic Server, e SPSS Modeler Server tenterà di elaborare la parte restante del flusso. Sarà necessario suddividere in sottocampioni i dataset di grosse dimensioni; ad esempio, collocando un nodo campione nel flusso.

Origine dati. Presupponendo che l'amministratore di SPSS Modeler Server abbia stabilito una connessione, selezionare l'origine dati che contiene i dati che si desidera utilizzare. Un'origine dati contiene i file e i metadati associati a quella origine. Fare clic su **Seleziona** per visualizzare un elenco delle origini dati disponibili. Per ulteriori informazioni, consultare l'argomento "Seleziona origine dati di Analytic Server".

Se è necessario creare una nuova origine dati o modificarne una esistente, fare clic su **Avvia editor origine dati...**

Seleziona origine dati di Analytic Server

La tabella Origini dati visualizza un elenco delle origini dati disponibili. Selezionare l'origine che si desidera utilizzare e fare clic su **OK**.

Fare clic su **Mostra proprietario** per visualizzare il proprietario dell'origine dati.

Filtra per consente di filtrare l'elenco di origini dati in base ad una **Parola chiave**, che verifica i criteri del filtro rispetto al nome e alla descrizione dell'origine dati, o in base al **Proprietario**. È possibile immettere

una combinazione di stringa, valore numerico o carattere jolly (%) come criteri del filtro. La stringa di ricerca è sensibile al maiuscolo/minuscolo. Fare clic su **Aggiorna** per aggiornare la tabella Origini dati.

Credenziali di Analytic Server

Se le proprie credenziali per accedere a Analytic Server sono diverse dalle credenziali per accedere a SPSS Modeler Server, è necessario immettere le credenziali di Analytic Server quando si esegue un flusso su Analytic Server. Se non si conoscono le credenziali, contattare l'amministratore del server.

Nodi supportati

Molti nodi SPSS Modeler sono supportati per l'esecuzione su HDFS, ma potrebbero esserci delle differenze nell'esecuzione di determinati nodi ed alcuni non sono attualmente supportati. Questo argomento descrive in dettaglio il livello corrente di supporto.

Generale

- Alcuni caratteri, che normalmente sono accettabili all'interno di un nome campo Modeler racchiuso tra virgolette, non verranno accettati da Analytic Server.
- Per poter essere eseguito in Analytic Server, un flusso Modeler deve iniziare con uno o più nodi origine Analytic Server e terminare in un unico nodo modellazione o nodo esportazione di Analytic Server; le unioni sono consentite ma non il forking.
- Si consiglia di impostare l'archiviazione dei target continui come reale piuttosto che intero. I modelli di calcolo del punteggio scrivono sempre valori reali nei file di dati di output per i target continui, mentre il modello di dati di output per i punteggi segue l'archiviazione del target. Quindi, se un target continuo ha un'archiviazione numero intero, si verificherà una mancata corrispondenza nei valori scritti e il modello dati per i punteggi e tale mancata corrispondenza causerà errori durante il tentativo di leggere i dati totalizzati.

Input

- Un flusso che inizia con qualcosa di diverso da un nodo origine Analytic Server verrà eseguito localmente.

Operazioni su record

Sono supportate tutte le operazioni su record. Di seguito sono riportate ulteriori note sulla funzionalità di tali nodi.

Seleziona

- Supporta lo stesso insieme di funzioni supportato dal nodo Ricava.

Campione

- Il campionamento a livello di blocco non è supportato.
- I metodi di campionamento complessi non sono supportati.

Aggregazione

- Le chiavi contigue non sono supportate.
- Le statistiche ordini (Mediana, Primo quartile, Terzo quartile) non sono supportate.

Ordina

- La scheda Ottimizzazione non è supportata.

In un ambiente distribuito, vi è un numero limitato di operazioni che mantiene l'ordine di record stabilito dal nodo Ordina.

- Un nodo Ordina seguito da un nodo di esportazione produce un'origine dati ordinata.
- Un nodo Ordina seguito da un nodo Campione con campionamento dei **Primi** record restituisce i primi *N* record.
- Un nodo Ordina seguito da un nodo di modellazione che ha come obiettivo **Ottimizza per dataset di grandi dimensioni** (Rete neurale, Lineare, C&R Tree, Quest o CHAID) è un modello utile per la redistribuzione casuale dei record tramite un ordinamento sulla

base di una chiave numero random derivata per evitare distorsioni che possono essere introdotte nell'algoritmo di creazione del modello se vengono ordinati i record originali.

In generale, collocare un nodo Ordina il più vicino possibile alle operazioni che necessitano dei record ordinati.

Unione

- L'unione in base all'ordine non è supportata.
- L'unione in base alla condizione non è supportata.
- La scheda Ottimizzazione non è supportata.
- Attualmente non è supportata la collocazione di un nodo Campione o di un nugget del modello tra un nodo Origine Analytic Server ed un nodo Unione. In genere è possibile specificare un nodo Seleziona per sostituire la funzionalità del nodo Campione.
- Analytic Server non esegue unioni con chiavi a stringhe vuote; vale a dire, se una delle chiavi in base a cui si sta eseguendo l'unione contiene stringhe vuote, i record che contengono la stringa vuota verranno esclusi dall'output dell'unione.
- Le operazioni di unione sono relativamente lente. Se si dispone di spazio sufficiente in HDFS, può essere molto più rapido unire le origini dati una volta sola ed utilizzare l'origine unita nei flussi successivi che unire le origini dati in ogni flusso.

Operazioni sui campi

Sono supportati i nodi Preparazione automatica dei dati, Tipo, Filtro, Ricava, Insieme, Riempimento, Ricodifica, Discretizza, Analisi RFM, Partizione, Crea indicatore, Riorganizza e Riordina campi. Di seguito sono riportate ulteriori note sulla funzionalità di tali nodi.

Preparazione automatica dei dati

- L'addestramento del nodo non è supportato. L'applicazione delle trasformazioni in un nodo Preparazione automatica dei dati addestrato a nuovi dati è supportata.

Tipo

- La colonna Controllo non è supportata.
- La scheda Formato non è supportata.

Ricava

- Tutte le funzioni di Ricava sono supportate, ad eccezione delle funzioni di sequenza.
- I campi di suddivisione non possono essere ricavati nello stesso flusso che li utilizza come suddivisioni; è necessario creare due flussi: uno che ricava il campo di suddivisione e uno che utilizza il campo come suddivisioni.
- Un campo indicatore non può essere utilizzato da solo in un confronto, ossia, `if (flagField) then ... endif` causerà un errore; la soluzione temporanea è utilizzare `if (flagField=trueValue) then ... endif`
- Quando si utilizza l'operatore `**` si consiglia di specificare l'esponente come numero reale, ad esempio `x**2.0`, invece di `x**2`, per ottenere la corrispondenza dei risultati in Modeler

Riempimento

- Supporta lo stesso insieme di funzioni supportato dal nodo Ricava.

Discretizza

La seguente funzionalità non è supportata.

- Discretizzazione ottimale
- Ranghi
- Blocchi -> Espansione: Somma dei valori
- Blocchi -> Pari merito: Mantieni in corrente e Assegna casualmente

- Blocchi ->N personalizzato: Valori oltre 100 e qualsiasi valore N dove 100 % N non è uguale a zero.

Analisi RFM

- L'opzione Mantieni in corrente per la gestione dei pari merito non è supportata. I punteggi monetari, l'attualità RFM e la frequenza non sempre corrispondono a quelli calcolati da Modeler dagli stessi dati. Gli intervalli di punteggio saranno gli stessi ma le assegnazioni del punteggio (numeri bin) potrebbero differire di uno.

Grafici

Tutti i nodi grafico sono supportati.

Modellazione

È supportato un numero limitato di nodi Modellazione, incluso: Lineare, Rete neurale, C&RT, Chaid, Quest. Di seguito sono riportate ulteriori note sulla funzionalità di tali nodi.

Lineare

Quando si creano modelli su dati di grosse dimensioni, in genere si modifica l'obiettivo su Dataset di grandi dimensioni o si specificano le suddivisioni.

- L'addestramento continuato dei modelli PSM esistenti non è supportato.
- L'obiettivo di creazione modello Standard è consigliato solo se sono definiti i campi di suddivisione in modo che il numero di record in ciascuna suddivisione non sia troppo elevato, dove la definizione di "troppo elevato" dipende dalla potenza dei singoli nodi nel cluster Hadoop. Di contro, è necessario anche assicurarsi che le suddivisioni non siano definite così finemente da essere presenti troppi pochi record per creare un modello.
- L'obiettivo Aumento non è supportato.
- L'obiettivo Bagging non è supportato.
- L'obiettivo Dataset di grandi dimensioni non è consigliato quando sono presenti pochi record; spesso non verrà creato alcun modello o ne verrà creato uno ridotto. Si potrebbero inoltre verificare problemi se i record di input sono ordinati in un modo sistematico che viola i presupposti di casualità degli algoritmi di creazione del modello dell'insieme utilizzati.
- La preparazione automatica dei dati non è supportata. Questo può causare problemi durante il tentativo di creazione di un modello di dati con numerosi valori mancanti; normalmente questi vengono assegnati come parte della preparazione automatica dei dati. Una soluzione temporanea potrebbe essere utilizzare un modello di struttura ad albero o una rete neurale con l'impostazione Avanzata per assegnare i valori mancanti selezionati.
- La statistica di accuratezza non viene calcolata per i modelli di suddivisione.

Rete neurale

Quando si creano modelli su dati di grosse dimensioni, in genere si modifica l'obiettivo su Dataset di grandi dimensioni o si specificano le suddivisioni.

- L'addestramento continuato dei modelli standard o PSM esistenti non è supportato.
- L'obiettivo di creazione modello Standard è consigliato solo se sono definiti i campi di suddivisione in modo che il numero di record in ciascuna suddivisione non sia troppo elevato, dove la definizione di "troppo elevato" dipende dalla potenza dei singoli nodi nel cluster Hadoop. Di contro, è necessario anche assicurarsi che le suddivisioni non siano definite così finemente da essere presenti troppi pochi record per creare un modello.
- L'obiettivo Aumento non è supportato.
- L'obiettivo Bagging non è supportato.
- L'obiettivo Dataset di grandi dimensioni non è consigliato quando sono presenti pochi record; spesso non verrà creato alcun modello o ne verrà creato uno ridotto. Si

potrebbero inoltre verificare problemi se i record di input sono ordinati in un modo sistematico che viola i presupposti di casualità degli algoritmi di creazione del modello dell'insieme utilizzati.

- Quando mancano numerosi valori nei dati, utilizzare l'impostazione Avanzata per assegnare i valori mancanti.
- La statistica di accuratezza non viene calcolata per i modelli di suddivisione.

C&R Tree, CHAID e Quest

Quando si creano modelli su dati di grosse dimensioni, in genere si modifica l'obiettivo su Dataset di grandi dimensioni o si specificano le suddivisioni.

- L'addestramento continuato dei modelli PSM esistenti non è supportato.
- L'obiettivo di creazione modello Standard è consigliato solo se sono definiti i campi di suddivisione in modo che il numero di record in ciascuna suddivisione non sia troppo elevato, dove la definizione di "troppo elevato" dipende dalla potenza dei singoli nodi nel cluster Hadoop. Di contro, è necessario anche assicurarsi che le suddivisioni non siano definite così finemente da essere presenti troppi pochi record per creare un modello.
- L'obiettivo Aumento non è supportato.
- L'obiettivo Bagging non è supportato.
- L'obiettivo Dataset di grandi dimensioni non è consigliato quando sono presenti pochi record; spesso non verrà creato alcun modello o ne verrà creato uno ridotto. Si potrebbero inoltre verificare problemi se i record di input sono ordinati in un modo sistematico che viola i presupposti di casualità degli algoritmi di creazione del modello dell'insieme utilizzati.
- Le sessioni interattive non sono supportate.
- La statistica di accuratezza non viene calcolata per i modelli di suddivisione.

Calcolo del punteggio del modello

I seguenti nugget del modello sono supportati per il calcolo del punteggio: C&RT, Quest, CHAID, Lineare, Regressione, Rete neurale, C5.0, Logistica, Genlin, GLMM, Cox, SVM, Rete di Bayes, TwoStep, KNN, Elenco di decisioni, Discriminante, Autoapprendimento, Rilevamento anomalie, Apriori, Carma, Medie K, Kohonen, R e Text Mining.

- Non verranno calcolate propensioni grezze o corrette. Come soluzione temporanea è possibile ottenere lo stesso effetto calcolando manualmente la propensione grezza utilizzando un nodo Ricava con la seguente espressione: `if 'valore-previsto' == 'valore-di-interesse' then 'prob-di-quel-valore' else 1-'prob-di-quel-valore' endif`
- Quando si calcola il punteggio di un modello, Analytic Server non controlla se tutti i campi utilizzati nel modello sono presenti nel dataset, quindi accertarsi che sia così prima dell'esecuzione in Analytic Server

R La sintassi R nel nugget deve essere costituita da operazioni record-at-a-time.

Output

I nodi Matrice, Analisi, Verifica dati, Trasformazioni, Statistiche e Medie sono supportati.

Esporta

Un flusso può iniziare con un nodo origine Analytic Server e terminare con un nodo di esportazione diverso dal nodo di esportazione di Analytic Server ma i dati si sposteranno da HDFS su SPSS Modeler Server, ed infine sulla posizione di esportazione.

Nodo origine IBM Cognos BI

Il nodo origine IBM Cognos BI consente di importare dati del database o report a elenco unico Cognos BI nella sessione di data mining. In questo modo è possibile combinare le funzioni di business intelligence di Cognos con le funzioni di analisi predittiva di IBM SPSS Modeler. È possibile importare dati relazionali, DMR (dimensionally-modeled relational) e OLAP.

Da una connessione a un server Cognos si seleziona innanzitutto un percorso da cui importare i dati o i report. Un percorso contiene un modello Cognos e tutte le cartelle, le query, i report, le visualizzazioni, i collegamenti, gli URL e le definizioni dei lavori associati al modello. Un modello Cognos definisce le regole di business, le descrizioni dei dati, le relazioni tra i dati, le dimensioni e le gerarchie di business e altre attività amministrative.

Se si stanno importando dei dati, a questo punto si devono scegliere gli oggetti da importare dal pacchetto selezionato. Gli oggetti importabili includono oggetti delle query (che rappresentano le tabelle di database) o singoli elementi delle query (che rappresentano le colonne delle tabelle). Per ulteriori informazioni, consultare l'argomento "Icane oggetti Cognos".

Se il pacchetto ha dei filtri definiti, è possibile importarne uno o più. Se un filtro che si importa è associato a dati importati, quel filtro viene applicato prima che i dati vengano importati. *Nota:* i dati da importare devono essere in formato UTF-8.













Se si sta importando un report, selezionare un pacchetto, oppure una cartella all'interno di un pacchetto, contenente uno o più report. Successivamente si seleziona il singolo report che si desidera importare. *Nota:* è possibile importare solo i report a elenco unico; gli elenchi multipli non sono supportati.

Se sono stati definiti dei parametri, per un oggetto di dati o per un report, è possibile specificarne i valori prima di importare l'oggetto o il report.

Icane oggetti Cognos

I vari tipi di oggetto che è possibile importare da un database Cognos BI sono rappresentati da diverse icone, come illustrato nella seguente tabella.

Tabella 3. Icone oggetti Cognos.

Icona	Oggetto
	Pacchetto
	Spazio dei nomi
	Oggetto della query
	Elemento della query
	Dimensione misure
	Misura
	Dimensione
	Gerarchia dei livelli
	Livello
	Filtro
	Report
	Calcolo standalone

Importazione di dati Cognos

Per importare i dati da un database IBM Cognos BI, nella scheda Dati della finestra di dialogo IBM Cognos BI verificare che l'opzione **Modalità** sia impostata su **Dati**.

Connessione. Fare clic sul pulsante **Modifica** per visualizzare una finestra di dialogo in cui è possibile definire i dettagli di una nuova connessione Cognos da cui importare dati o report. Se si è già connessi a un server Cognos mediante IBM SPSS Modeler è possibile anche modificare i dettagli della connessione corrente. Per ulteriori informazioni, consultare l'argomento "Connessioni Cognos" a pagina 37.

Posizione. Una volta stabilita la connessione al server Cognos, fare clic sul pulsante **Modifica** accanto a questo campo per visualizzare un elenco dei package disponibili da cui importare il contenuto. Per ulteriori informazioni, consultare l'argomento "Selezione del percorso Cognos" a pagina 37.

Contenuto. Visualizza il nome del package selezionato, insieme agli spazi dei nomi associati al package. Fare doppio clic su uno spazio dei nomi per visualizzare gli oggetti che è possibile importare. I vari tipi di oggetti sono indicati da icone diverse. Per ulteriori informazioni, consultare l'argomento "Icône oggetti Cognos" a pagina 35.

Per scegliere un oggetto da importare, selezionare l'oggetto e fare clic sulla freccia più in alto delle due frecce a destra per spostare l'oggetto nel riquadro **Campi da importare**. Se si seleziona l'oggetto di una query, vengono importati tutti gli elementi della query. Fare doppio clic sull'oggetto di una query per espanderla in modo da poter scegliere uno o più dei suoi elementi. Per selezionare più elementi, utilizzare la combinazione Ctrl-clic (selezione di singoli elementi), Maiusc-clic (selezione di un blocco di elementi) e Ctrl-A (selezione di tutti gli elementi).

Per scegliere un filtro da applicare (se nel pacchetto sono stati definiti dei filtri), spostarsi sul filtro nel riquadro **Contenuto**, selezionare il filtro e fare clic sulla freccia più in basso delle due frecce a destra, per spostare il filtro nel riquadro **Filtri da applicare**. È possibile eseguire più selezioni utilizzando la combinazione Ctrl-clic del mouse (per selezionare singoli filtri) e Maiusc-clic del mouse (per selezionare un gruppo di filtri).

Campi da importare. Elenca gli oggetti del database selezionati per essere importati in IBM SPSS Modeler per l'elaborazione. Se un determinato oggetto non è più necessario, selezionarlo e fare clic sulla freccia sinistra per inserirlo nuovamente nel riquadro **Contenuto**. Per selezionare più filtri, utilizzare le stesse combinazioni di tasti descritte per **Contenuto**.

Filtri da applicare. Elenca i filtri che si è scelto di applicare ai dati prima della loro importazione. Se un determinato filtro non è più necessario, selezionarlo e fare clic sulla freccia sinistra per inserirlo nuovamente nel riquadro **Contenuto**. Per selezionare più filtri, utilizzare le stesse combinazioni di tasti descritte per **Contenuto**.

Parametri. Se questo pulsante è abilitato significa che l'oggetto selezionato ha dei parametri definiti. È possibile utilizzare i parametri per apportare delle modifiche (per esempio eseguire un calcolo parametrizzato) prima di importare i dati. Se i parametri sono definiti ma non è fornito alcun valore di default, il pulsante visualizza un triangolo di avviso. Fare clic sul pulsante per visualizzare i parametri e modificarli se necessario. Se il pulsante è disabilitato significa che il report non ha parametri definiti.

Aggrega dati prima dell'importazione. Selezionare questa casella se si desidera importare i dati aggregati anziché i dati grezzi.

Importazione di report Cognos

Per importare un report predefinito da un database IBM Cognos BI, nella scheda Dati della finestra di dialogo IBM Cognos BI, verificare che l'opzione **Modalità** sia impostata su **Report**. *Nota:* è possibile importare solo i report a elenco unico; gli elenchi multipli non sono supportati.

Connessione. Fare clic sul pulsante **Modifica** per visualizzare una finestra di dialogo in cui è possibile definire i dettagli di una nuova connessione Cognos da cui importare dati o report. Se si è già connessi a un server Cognos mediante IBM SPSS Modeler è possibile anche modificare i dettagli della connessione corrente. Per ulteriori informazioni, consultare l'argomento "Connessioni Cognos".

Posizione. Una volta stabilita la connessione al server Cognos, fare clic sul pulsante **Modifica** accanto a questo campo per visualizzare un elenco dei package disponibili da cui importare il contenuto. Per ulteriori informazioni, consultare l'argomento "Selezione del percorso Cognos".

Contenuto. Visualizza il nome del pacchetto selezionato o della cartella che contiene i report. Andare al report specifico, selezionarlo e fare clic sulla freccia destra per portare il report nel campo **Report da importare**.

Report da importare. Indica il report che si è scelto di importare in IBM SPSS Modeler. Se il report non è più necessario, selezionarlo e fare clic sulla freccia sinistra per inserirlo nuovamente nel riquadro **Contenuto**, oppure portare un report diverso nel campo.

Parametri. Se questo pulsante è abilitato significa che il report selezionato ha dei parametri definiti. È possibile utilizzare i parametri per apportare delle modifiche prima di importare il report (per esempio, specificare una data di inizio e di fine per i dati del report). Se i parametri sono definiti ma non è fornito alcun valore di default, il pulsante visualizza un triangolo di avviso. Fare clic sul pulsante per visualizzare i parametri e modificarli se necessario. Se il pulsante è disabilitato significa che il report non ha parametri definiti.

Connessioni Cognos

La finestra di dialogo Connessioni Cognos consente di selezionare il server Cognos BI da utilizzare per l'importazione o l'esportazione degli oggetti del database.

URL server Cognos. Immettere l'URL del server Cognos BI da cui importare o esportare. Si tratta del valore della proprietà di ambiente "URI dispatcher esterno" di IBM Cognos Configuration sul server Cognos BI. Se non si sa quale URL utilizzare, rivolgersi all'amministratore del sistema Cognos.

Modalità. Selezionare **Imposta credenziali** se si desidera accedere con uno spazio dei nomi, un nome utente ed una password Cognos specifici (ad esempio, come amministratore). Selezionare **Utilizza connessione Anonima** per accedere senza credenziali utente, nel qual caso gli altri campi non vengono compilati.

Spazio dei nomi. Specificare il provider di protezione per l'autenticazione Cognos utilizzato per accedere al server. Il provider per l'autenticazione si utilizza per definire e gestire utenti, gruppi e ruoli e per controllare il processo di autenticazione.

Nome utente. Immettere il nome utente Cognos con cui accedere al server.

Password. Immettere la password associata al nome utente specificato.

Salva come default. Fare clic su questo pulsante per memorizzare i valori come impostazioni di default ed evitare di doverli inserire nuovamente ogni volta che si apre il nodo.

Selezione del percorso Cognos

La finestra di dialogo Specifica percorso consente di selezionare un pacchetto Cognos da cui importare i dati, oppure un pacchetto o una cartella da cui importare i report.

Cartelle pubbliche. Se si stanno importando dei dati, questa opzione elenca i pacchetti e le cartelle disponibili dal server scelto. Selezionare il pacchetto da utilizzare e fare clic su **OK**. Per ogni nodo origine Cognos BI è possibile selezionare un solo pacchetto.

Se si stanno importando dei report, questa opzione elenca le cartelle e i pacchetti contenenti report che sono disponibili dal server scelto. Selezionare un pacchetto o una cartella di report e fare clic su **OK**. È possibile selezionare solo un pacchetto o una cartella di report per ogni nodo origine Cognos BI, benché le cartelle di report possano contenere altre cartelle di report, oppure singoli report.

Specifica dei parametri per i dati o i report

Se sono stati definiti dei parametri in Cognos BI, per un oggetto di dati o per un report, è possibile specificarne i valori prima di importare l'oggetto o il report. Un esempio di parametri per un report sarebbero le date di inizio e di fine dei contenuti del report.

Nome. Il nome del parametro come specificato nel database Cognos BI.

Tipo. Descrizione del parametro.

Valore. Il valore da assegnare al parametro. Per specificare o modificare un valore, fare doppio clic sulla relativa cella nella tabella. I valori non vengono convalidati qui, pertanto eventuali valori non validi saranno rilevati al momento dell'esecuzione.

Rimuovi automaticamente i parametri non validi dalla tabella. Questa opzione è selezionata di default e rimuove tutti i parametri non validi rilevati nell'oggetto di dati o nel report.

Nodo origine IBM Cognos TM1

Il nodo origine IBM Cognos TM1 consente di portare i dati Cognos TM1 nella sessione di data mining. In questo modo è possibile combinare le funzioni di pianificazione enterprise di Cognos con le funzioni di analisi predittiva di IBM SPSS Modeler. È possibile importare una versione meno complessa dei dati del cubo OLAP multidimensionale.

È necessario modificare i dati in TM1 prima di importarli. *Nota:* i dati da importare devono essere in formato UTF-8.

Da una connessione a un hub PM Cognos, selezionare prima un server TM1 da cui importare i dati; un server contiene uno o più cubi TM1. Selezionare quindi il cubo richiesto e all'interno del cubo selezionare le colonne e le righe che si desidera importare.

Nota: Prima di poter utilizzare i nodi origine o di esportazione di TM1 in SPSS Modeler, è necessario copiare i seguenti tre processi dal Modeler al server TM1: *ExportToSPSS.pro*, *ImportFromSPSS.pro* e *SPSSCreateNewMeasures.pro*. Per aggiungere questi processi al server TM1 è necessario copiare i file nella directory dei dati del server TM1 e riavviare il server TM1. Questi file sono disponibili nella seguente directory: <Directory di installazione Modeler>/ext/bin/pasw.tm1/scripts.

Importazione di dati Cognos TM1

Per importare i dati da un database IBM Cognos TM1, sulla scheda Dati della finestra di dialogo IBM Cognos TM1 selezionare l'hub PM TM1 rilevante e il server, il cubo e i dettagli dei dati associati.

Nota: Prima di importare i dati è necessario effettuare alcune operazioni di preelaborazione in TM1 per assicurarsi che i dati siano in un formato riconoscibile a IBM SPSS Modeler. Questo implica filtrare i dati utilizzando Subset Editor per ottenere una vista con dimensione e forma corrette per l'importazione.

Sistema PM. Immettere l'URL dell'hub che contiene il server TM1 a cui ci si desidera **Connettere**.

Server TM1. Dopo aver stabilito la connessione all'hub Cognos, selezionare il server che contiene i dati da importare e fare clic su **Login**. Se non ci si è mai connessi in precedenza a questo server viene richiesta l'immissione del **Nome utente** e della **Password**; in alternativa, è possibile selezionare un altro server.

Selezionare una vista cubo TM1 da importare. Visualizza il nome dei cubi all'interno del server TM1 da cui è possibile importare i dati. Fare doppio clic su un cubo per visualizzare i dati della vista che è possibile importare.

Per scegliere i dati da importare, selezionare la vista e fare clic sulla freccia destra per spostarla nel pannello **Vista da importare**. Se la vista necessaria non è visibile, fare doppio clic su un cubo per espanderne l'elenco delle viste.

Dimensione colonna. Elenca il nome della dimensione colonna nei dati che si è scelto di importare. Scorrere l'elenco di livelli e selezionare quello richiesto.

Dimensioni riga. Elenca il nome della dimensione riga nei dati che si è scelto di importare. Scorrere l'elenco di livelli e selezionare quello richiesto.

Dimensioni contesto. Visualizza solo. Visualizza le dimensioni del contesto correlate alle colonne e alle righe selezionate.

Nodo origine SAS

Nota: questa funzione è disponibile in SPSS Modeler Professional e SPSS Modeler Premium.

Il nodo origine SAS consente di importare dati in formato SAS nella sessione di data mining. È possibile importare quattro tipi di file:

- SAS per Windows/OS2 (.sd2)
- SAS per UNIX (.ssd)
- File di trasporto SAS (.tpt)
- SAS Versione 7/8/9 (.sas7bdat)

Quando i dati vengono importati, tutte le variabili vengono mantenute e non viene modificato alcun tipo di variabile. Vengono selezionati tutti i casi.

Impostazione delle opzioni relative al nodo origine SAS

Importa. Selezionare il tipo di file SAS da trasportare. È possibile selezionare **SAS per Windows/OS2 (.sd2)**, **SAS per UNIX (.SSD)**, **File di trasporto SAS (.tpt)** o **SAS Versione 7/8/9 (.sas7bdat)**.

File di importazione. Specificare il nome del file. È possibile immettere un nome di file o fare clic sul pulsante con i puntini di sospensione (...) per passare alla posizione del file.

Membro. Selezionare un membro da importare dal file di trasporto SAS selezionato in precedenza. È possibile immettere un nome per il membro oppure fare clic su **Seleziona** per scorrere l'elenco di tutti i membri nel file.

Leggi formati utente da file di dati SAS. Selezionare per leggere i formati utente. I file SAS archiviano dati e formati di dati (per esempio etichette di variabile) in file diversi. Nella maggior parte dei casi, sarà necessario importare anche i formati. Se invece si utilizza un insieme di dati di grandi dimensioni, può essere utile deselezionare questa opzione per risparmiare memoria.

File formato. Se è richiesto un file di formato, questa casella di testo è attivata. È possibile immettere un nome di file o fare clic sul pulsante con i puntini di sospensione (...) per passare alla posizione del file.

Nomi delle variabili. Selezionare un metodo per la gestione di etichette e nomi di variabili dopo l'importazione da un file SAS. I metadati che si sceglie di includere vengono mantenuti per la sessione di IBM SPSS Modeler e possono essere esportati nuovamente per essere utilizzati in SAS.

- **Leggi nomi ed etichette.** Selezionare per leggere le etichette ed i nomi di variabili in IBM SPSS Modeler. Questa opzione è selezionata per default e i nomi di variabile vengono visualizzati nel nodo Tipo. Le etichette possono essere visualizzate nel generatore di espressioni, nei grafici, nei browser dei modelli e in altri tipi di output, in base alle opzioni specificate nella finestra di dialogo delle proprietà del flusso.
- **Leggi etichette come nomi.** Selezionare per leggere le etichette di variabile descrittive dal file SAS piuttosto che i nomi di campo brevi ed utilizzare tali etichette come nomi di variabili in IBM SPSS Modeler.

Nodo origine Excel

Il nodo origine Excel consente di importare dati da qualsiasi versione di Microsoft Excel.

Tipo di file. Selezionare il tipo di file Excel che si sta importando.

File di importazione. Specifica il nome e la posizione del file del foglio di calcolo da importare.

Utilizza intervallo denominato. Consente di specificare un intervallo denominato di celle, come definito nel foglio di lavoro Excel. Fare clic sul pulsante con i puntini di sospensione (...) per scegliere dall'elenco di intervalli disponibili. Se viene utilizzato un intervallo denominato, le altre impostazioni del foglio di lavoro e dell'intervallo di dati non sono più applicabili e vengono pertanto disattivate.

Scegli foglio di lavoro. Specifica il foglio di lavoro da importare, in base all'indice o al nome.

- **Tramite indice.** Specifica il valore dell'indice del foglio di lavoro che si desidera importare, iniziando con 0 per il primo foglio di lavoro, 1 per il secondo e così via.
- **Per nome.** Specificare il nome del foglio di lavoro che si desidera importare. Fare clic sul pulsante con i puntini di sospensione (...) per scegliere dall'elenco di fogli di lavoro disponibili.

Intervallo su foglio di lavoro. È possibile importare i dati iniziando dalla prima riga non vuota o da un intervallo esplicito di celle.

- **L'intervallo inizia alla prima riga non vuota.** Individua la prima cella non vuota e la utilizza come angolo superiore sinistro dell'intervallo di dati.
- **Intervallo esplicito di celle.** Consente di specificare un intervallo esplicito in base alla riga e colonna. Per esempio, per specificare l'intervallo Excel A1:D5, è possibile inserire A1 nel primo campo e D5 nel secondo (o in alternativa, R1C1 e R5C4). Vengono restituite tutte le righe dell'intervallo specificato, incluse quelle vuote.

Su righe vuote. Se si incontra più di una riga vuota, è possibile decidere se **interrompere la lettura** oppure selezionare **Restituisci righe vuote** per continuare a leggere tutti i dati fino alla fine del foglio di lavoro, incluse le righe vuote.

La prima riga ha nomi di colonna. Indica che la prima riga dell'intervallo specificato deve essere utilizzata come nome di campo (colonna). Se l'opzione non è selezionata, i nomi di campo vengono generati automaticamente.

Archiviazione e livello di misurazione dei campi

Quando si leggono i valori da Excel, per default i campi con archiviazione numerica vengono letti con un livello di misurazione *Continuo* e i campi stringa come *Nominali*. Nella scheda Tipo è possibile modificare manualmente il livello di misurazione (continuo o nominale), ma l'archiviazione viene determinata automaticamente (tuttavia, se necessario, è possibile modificarla utilizzando una funzione di conversione, quale `to_integer`, in un nodo Riempimento o Ricava). Per ulteriori informazioni, consultare l'argomento "Impostazione dell'archiviazione e della formattazione dei campi" a pagina 24.

Per default, i campi contenenti sia valori numerici che valori stringa vengono letti come numeri, cioè i valori stringa vengono impostati su valori null (valori mancanti di sistema) in IBM SPSS Modeler. Ciò si verifica perché — a differenza di Excel — IBM SPSS Modeler non consente l'utilizzo di tipi di archiviazione misti all'interno di un campo. Per evitare questa situazione, è possibile impostare il formato delle celle su **Testo** nel foglio di lavoro Excel, in modo che tutti i valori, inclusi i numeri, vengano letti come stringhe.

Nodo origine XML

Nota: questa funzione è disponibile in SPSS Modeler Professional e SPSS Modeler Premium.

Il nodo origine XML consente di importare i dati da un file in formato XML in un flusso IBM SPSS Modeler. XML è un linguaggio standard per lo scambio di dati e per molte organizzazioni rappresenta il formato preferito per questo scopo. Ad esempio, l'agenzia delle entrate desidera analizzare i dati delle entrate fiscali inviati online ed in formato XML (consultare <http://www.w3.org/standards/xml/>).

L'importazione dei dati XML in un flusso IBM SPSS Modeler consente di eseguire una vasta gamma di funzioni di analisi predittive sui dati di input. I dati XML vengono analizzati in un formato tabulare nel quale le colonne corrispondono ai diversi livelli di nidificazione degli elementi e degli attributi XML. Gli elementi XML vengono visualizzati in formato XPath (vedere <http://www.w3.org/TR/xpath20/>).

Leggi un singolo file. Per default, IBM SPSS Modeler legge un singolo file, specificato nel campo **Sorgente dati XML**.

Leggi tutti i file XML in una directory. Scegliere questa opzione se si desidera leggere tutti i file XML in una particolare directory. Specificare la posizione nel campo **Directory** visualizzato. Selezionare la casella di controllo **Includi sottodirectory** per leggere anche i file XML di tutte le sottodirectory della directory specificata.

Sorgente dati XML. Immettere il percorso e il nome file completi del file di origine XML che si desidera importare, oppure utilizzare il pulsante **Visualizza** per trovare il file.

Schema XML. (Opzionale) Specificare il percorso e il nome file completi di un file XSD o DTD dal quale si desidera leggere la struttura XML, oppure utilizzare il pulsante **Visualizza** per trovare questo file. Se si lascia vuoto questo campo, la struttura viene letta dal file di origine XML. Un file XSD o DTD può includere più elementi root. In tal caso, quando viene selezionato un campo diverso, viene visualizzata una finestra di dialogo nella quale è possibile selezionare l'elemento root che si desidera utilizzare. Per ulteriori informazioni, consultare l'argomento "Selezione da più elementi root" a pagina 42.

Struttura XML. Una struttura ad albero gerarchica che mostra la struttura del file di origine XML (o dello schema, se specificato nel campo **Schema XML**). Per definire un limite del record, selezionare un elemento e fare clic sul pulsante con la freccia destra per copiarlo nel campo **Record**.

Visualizza attributi. Visualizza o nasconde gli attributi degli elementi XML nel campo **Struttura XML**.

Record (espressione XPath). Mostra la sintassi XPath per un elemento copiato dal campo della struttura XML. Questo elemento viene quindi evidenziato nella struttura XML e definisce il limite del record. Ogni volta che si incontra questo elemento nel file di origine, viene creato un nuovo record. Se questo campo è vuoto, come limite del record viene utilizzato il primo elemento figlio nella radice.

Leggi tutti i dati. Per default, tutti i dati nel file di origine vengono letti nel flusso.

Specifica i dati da leggere. Selezionare questa opzione se si desidera importare singoli elementi o attributi o entrambi. Selezionando questa opzione si attiva la tabella **Campi**, nella quale è possibile specificare i dati che si desidera importare.

Campi. Questa tabella elenca gli elementi e gli attributi selezionati per l'importazione, se è stata selezionata l'opzione **Specifica i dati da leggere**. È possibile immettere la sintassi XPath di un elemento o di un attributo direttamente nella colonna XPath, oppure selezionare un elemento o un attributo nella struttura XML e fare clic sul pulsante con la freccia destra per copiare l'elemento nella tabella. Per copiare tutti gli elementi e gli attributi figlio di un elemento, selezionare l'elemento nella struttura XML e fare clic sul pulsante a doppia freccia.

- **XPath.** La sintassi XPath degli elementi da importare.
- **Posizione.** La posizione nella struttura XML degli elementi da importare. **Percorso fisso** mostra il percorso dell'elemento relativo all'elemento evidenziato nella struttura XML (o al primo elemento figlio nella radice se non è evidenziato alcun elemento). **Qualsiasi posizione** indica un elemento con il nome dato in qualsiasi posizione nella struttura XML. **Personalizzato** viene visualizzato se si immette una posizione direttamente nella colonna XPath.

Selezione da più elementi root

A differenza dei file XML che possono includere un unico elemento root, i file XSD o DTD possono contenere più elementi radice. Se presente, viene utilizzata la root corrispondente alla root del file di origine XML. In caso contrario, è necessario selezionare la root da utilizzare.

Selezionare la radice da visualizzare. Selezionare l'elemento root che si desidera utilizzare. La root di default corrisponde al primo elemento root della struttura XSD o DTD.

Rimozione degli spazi indesiderati dai dati di origine XML

È possibile implementare le interruzioni di linea nei dati di origine XML utilizzando la combinazione di caratteri [CR][LF]. In alcuni casi, tali interruzioni di linea possono essere inserite al centro di una stringa di testo, ad esempio:

```
<description>An in-depth look at creating applications[CR][LF]
with XML.</description>
```

Le interruzioni di linea potrebbero non essere visibili quando il file viene aperto in alcune applicazioni, ad esempio un browser Web. Tuttavia, quando i dati vengono letti nel flusso attraverso il nodo origine XML, le interruzioni di linea vengono convertite in una serie di caratteri spazio.

Il problema può essere risolto utilizzando un nodo Riempimento per la rimozione degli spazi indesiderati:

Di seguito è riportato un esempio di come ottenere il risultato desiderato:

1. Collegare un nodo Riempimento al nodo origine XML.
2. Aprire il nodo Riempimento e utilizzare il selettore di campo per selezionare il campo con gli spazi indesiderati.
3. Impostare la condizione **Sostituisci** su **In base alla condizione** e impostare **Condizione** su **vera**.
4. Nel campo **Sostituisci con**, immettere `replace(" ", "", @FIELD)` e fare clic su OK.
5. Collegare un nodo Tabella al nodo Riempimento ed eseguire il flusso.

Nell'output del noto Tabella, il testo viene ora visualizzato senza spazi aggiuntivi.

Nodo Genera simulazione

Il nodo Genera simulazione fornisce un modo semplice per generare dati simulati, senza dati cronologici utilizzando distribuzioni statistiche specificate dagli utenti o automaticamente utilizzando le distribuzioni ottenute dall'esecuzione di un nodo Adattamento simulazione sui dati cronologici esistenti. La generazione di dati simulati è utile quando si desidera valutare il risultato di un modello predittivo in presenza di incertezza negli input del modello.

Creazione di dati senza dati cronologici

Il nodo Genera simulazione è disponibile dalla tavolozza Origini e può essere aggiunto direttamente all'area di disegno del flusso.

1. Fare clic sulla scheda **Input** della palette dei nodi.
2. Trascinare e rilasciare oppure fare doppio clic per aggiungere il nodo Genera simulazione all'area di disegno del flusso.
3. Fare doppio clic per aprire la relativa finestra di dialogo e specificare i campi, i tipi di archiviazione, le distribuzioni statistiche e i parametri di distribuzione.

Nota: i nodi Genera simulazione selezionati dalla tavolozza Origini saranno completamente vuoti, senza campi e senza informazioni sulla distribuzione. Questo consente di poter creare interamente i dati simulati senza dati cronologici.

Generazione di dati simulati utilizzando i dati cronologici esistenti

Un nodo Genera simulazione può inoltre essere creato eseguendo un nodo terminale Adattamento simulazione:

1. Fare clic con il tasto destro del mouse sul nodo Adattamento simulazione e scegliere **Esegui** dal menu.
2. Il nodo Genera simulazione è presente nell'area di disegno del flusso con un collegamento di aggiornamento al nodo Adattamento simulazione.
3. Quando viene generato, il nodo Genera simulazione eredita tutti i campi, i tipi di archiviazione e le informazioni statistiche della distribuzione dal nodo Adattamento simulazione.

Definizione di un collegamento di aggiornamento ad un nodo Adattamento simulazione

È possibile creare un collegamento tra un nodo Genera simulazione e un nodo Adattamento simulazione. È utile se si desidera aggiornare uno o più campi con le informazioni relative alla distribuzione con miglior adattamento, determinate dall'adattamento ai dati cronologici.

1. Fare clic con il tasto destro del mouse sul nodo Genera simulazione.
2. Dal menu, selezionare **Definisci collegamento di aggiornamento**. Il cursore si trasforma in cursore del collegamento.
3. Fare clic su un altro nodo. Se questo nodo è un nodo Adattamento simulazione, viene creato un collegamento. Se il nodo non è un nodo Adattamento simulazione, non viene creato un collegamento e il cursore torna ad essere un cursore normale.

Se i campi nel nodo Adattamento simulazione sono diversi da quelli del nodo Genera simulazione, viene visualizzato un messaggio che informa l'utente che vi sono differenze.

Quando viene utilizzato il nodo Adattamento simulazione per aggiornare il nodo Genera simulazione collegato, il risultato dipende dalla presenza o meno degli stessi campi in entrambi i nodi e se i campi sono sbloccati nel nodo Genera simulazione. I risultati dell'aggiornamento del nodo Adattamento simulazione vengono visualizzati nella tabella riportata di seguito.

Tabella 4. Risultati dell'aggiornamento di un nodo Adattamento simulazione

Campo nel nodo Genera simulazione	Campo nella simulazione Nodo Adattamento	
	Presente	Mancante
Presente e sbloccato.	Il campo viene sovrascritto.	Il campo viene eliminato.
Mancante/i.	Il campo viene aggiunto.	Nessuna modifica.

Tabella 4. Risultati dell'aggiornamento di un nodo Adattamento simulazione (Continua)

	Campo nella simulazione	Nodo Adattamento
Presente e bloccato.	La distribuzione del campo non viene sovrascritta. Le informazioni nella finestra di dialogo Dettagli adattamento e le correlazioni vengono aggiornate.	Il campo non viene sovrascritto. Le correlazioni sono impostate su zero.
La casella di controllo Non cancellare Min/Max durante il riadattamento è selezionata.	Il campo viene sovrascritto, tranne i valori della colonna Min, Max.	
La casella di controllo Non ricalcolare le correlazioni durante il riadattamento è selezionata.	Se il campo è sbloccato, viene sovrascritto.	Le correlazioni non vengono sovrascritte.

Rimozione di un collegamento di aggiornamento ad un nodo Adattamento simulazione

È possibile rimuovere un collegamento tra un nodo Genera simulazione ed un nodo Adattamento simulazione effettuando le seguenti operazioni:

1. Fare clic con il tasto destro del mouse sul nodo Genera simulazione.
2. Dal menu, selezionare **Rimuovi collegamento aggiornamento**. Il collegamento viene rimosso.

Impostazione delle opzioni relative al nodo Genera simulazione

È possibile utilizzare le opzioni presenti nella scheda Dati della finestra di dialogo del nodo Genera simulazione per effettuare le seguenti operazioni:

- Visualizzare, specificare e modificare le informazioni sulla distribuzione statistica per i campi.
- Visualizzare, specificare e modificare le correlazioni tra i campi.
- Specificare il numero di iterazioni e casi da simulare.

Seleziona un elemento. Consente di cambiare vista tra le tre disponibili per il nodo Genera simulazione: Campi simulati, Correlazioni e Opzioni avanzate.

Vista Campi simulati

Se il nodo Genera simulazione è stato generato o aggiornato da un nodo Adattamento simulazione utilizzando dati cronologici, nella vista Campi simulati è possibile visualizzare e modificare le informazioni sulla distribuzione statistica per ciascun campo. Le seguenti informazioni relative a ciascun campo vengono copiate nella scheda **Tipi** del nodo Genera simulazione dal nodo Adattamento simulazione:

- Livello di misurazione
- Valori
- Mancante
- Controllo
- Ruolo

Se non si dispone di dati cronologici, è possibile definire i campi e specificare le relative distribuzioni selezionando un tipo di archiviazione, e selezionando un tipo di distribuzione ed immettendo i parametri richiesti. La generazione dei dati secondo questo metodo implica che le informazioni relative al livello di misurazione di ciascun campo non saranno disponibili fino a quando non verranno create istanze dei dati, ad esempio sulla scheda **Tipi** o in un nodo Tipo.

La vista Campi simulati contiene alcuni strumenti, che è possibile utilizzare per eseguire le seguenti attività:

- Aggiungere e rimuovere campi.
- Modificare l'ordine dei campi visualizzati.
- Specificare un tipo di archiviazione per ciascun campo.
- Specificare una distribuzione statistica per ogni campo.
- Specificare i valori dei parametri per la distribuzione statistica di ciascun campo.

Campi simulati. Questa tabella contiene una riga vuota se il nodo Genera simulazione è stato aggiunto a un'area di disegno del flusso dalla tavolozza Origini. Quando questa riga viene modificata, viene aggiunta una nuova riga vuota in fondo alla tabella. Se il nodo Genera simulazione è stato creato da un nodo Adattamento simulazione, questa tabella conterrà una riga per ogni campo di dati cronologici. È possibile aggiungere ulteriori righe alla tabella facendo clic sull'icona **Aggiungi nuovo campo**.

La tabella Campi simulati è composta dalle seguenti colonne:

- **Campo.** Contiene i nomi dei campi. I nomi dei campi possono essere modificati immettendo i dati nelle celle.
- **Archiviazione.** Le celle di questa colonna contengono un elenco a discesa dei tipi di archiviazione. I tipi di archiviazione disponibili sono **Stringa, Intero, Reale, Ora, Data e Timestamp**. La scelta del tipo di archiviazione determina quali distribuzioni sono disponibili nella colonna Distribuzione. Se il nodo Genera simulazione è stato creato da un nodo Adattamento simulazione, il tipo di archiviazione viene copiato dal nodo Adattamento simulazione.

Nota: per i campi con tipo di memorizzazione data/ora, è necessario specificare i parametri di distribuzione come numeri interi. Ad esempio, per specificare la data media 1 gennaio 1970, utilizzare il numero intero 0. Il numero intero firmato rappresenta il numero di secondi a partire da (o prima di) mezzanotte del 1 gennaio 1970.

- **Stato.** Le icone nella colonna Stato indicano lo stato adattamento di ciascun campo.



Non è stata specificata una distribuzione per il campo o mancano uno o più parametri di distribuzione. Per poter eseguire la simulazione, è necessario specificare una distribuzione per questo campo ed immettere valori validi per i parametri.



Il campo è impostato sulla distribuzione più adatta. **Nota:** questa icona può essere visualizzata solo se il nodo Genera simulazione viene creato da un nodo Adattamento simulazione.



La distribuzione più adatta è stata sostituita con una distribuzione alternativa dalla finestra di dialogo secondaria Dettagli adattamento. Per ulteriori informazioni, consultare l'argomento "Dettagli adattamento" a pagina 50.



La distribuzione è stata specificata o modificata manualmente, e potrebbe includere un parametro specificato in più di un livello.

- **Bloccato.** Il blocco di un campo simulato, mediante la selezione della casella di controllo nella colonna con l'icona del blocco, esclude il campo dall'aggiornamento automatico da parte di un nodo Adattamento simulazione collegato. Ciò è particolarmente utile quando si specifica manualmente una

distribuzione e si desidera assicurarsi che non sarà interessata dall'adattamento distribuzione automatico quando viene eseguito un nodo Adattamento simulazione collegato.

- **Distribuzione.** Le celle di questa colonna contengono un elenco a discesa delle distribuzioni statistiche. La scelta del tipo di archiviazione determina quali distribuzioni sono disponibili in questa colonna per un determinato campo. Per ulteriori informazioni, consultare l'argomento "Distribuzioni" a pagina 53.

Nota: Non è possibile specificare la distribuzione Fissa per ogni campo. Se si desidera che ogni campo nei dati generati sia fisso, è possibile utilizzare un nodo di input utente seguito da un nodo di bilanciamento.

- **Parametri.** I parametri di distribuzione associati a ciascuna distribuzione adattata sono visualizzati in questa colonna. I valori multipli di un parametro sono separati da virgole. La specifica di più valori per un parametro genera più iterazioni per la simulazione. Per ulteriori informazioni, consultare l'argomento "Iterazioni" a pagina 52. Se mancano i parametri, questo si riflette nell'icona visualizzata nella colonna Stato. Per specificare i valori per i parametri, fare clic su questa colonna nella riga corrispondente al campo di interesse e scegliere **Specifica** dall'elenco. In questo modo viene aperta la finestra di dialogo secondaria Specifica parametri. Per ulteriori informazioni, consultare l'argomento "Specifica parametri" a pagina 51. Questa colonna è disabilitata se viene scelto Empirica nella colonna Distribuzione.
- **Min, Max.** In questa colonna, per alcune distribuzioni, è possibile specificare un valore minimo, un valore massimo o entrambi per i dati simulati. I dati simulati inferiori al valore minimo e superiori al valore massimo verranno rifiutati, anche se sarebbero validi per la distribuzione specificata. Per specificare i valori minimo e massimo, fare clic su questa colonna nella riga corrispondente al campo di interesse e scegliere **Specifica** dall'elenco. In questo modo viene aperta la finestra di dialogo secondaria Specifica parametri. Per ulteriori informazioni, consultare l'argomento "Specifica parametri" a pagina 51. Questa colonna è disabilitata se viene scelto Empirica nella colonna Distribuzione.

Utilizza adattamento migliore. È abilitato solo se il nodo Genera simulazione è stato creato automaticamente da un nodo Adattamento simulazione utilizzando dati cronologici, ed è selezionata una singola riga nella tabella Campi simulati. Sostituisce le informazioni del campo nella riga selezionata con le informazioni della distribuzione più adatta per il campo. Se le informazioni nella riga selezionata sono state modificate, premendo questo pulsante le informazioni verranno reimpostate sulla distribuzione più adatta determinata dal nodo Adattamento simulazione.

Dettagli adattamento. È abilitato solo se il nodo Genera simulazione è stato creato automaticamente da un nodo Adattamento simulazione. Consente di aprire la finestra di dialogo secondaria Dettagli adattamento. Per ulteriori informazioni, consultare l'argomento "Dettagli adattamento" a pagina 50.

Diverse attività utili possono essere eseguite utilizzando le icone a destra della vista Campi simulati. Le icone sono descritte nella tabella seguente.

Tabella 5. Icone nella vista Campi simulati









Icona	Suggerimento	Descrizione
	Modifica parametri di distribuzione	È abilitato solo quando è selezionata una singola riga nella tabella Campi simulati. Consente di aprire la finestra di dialogo secondaria Specifica parametri per la riga selezionata. Per ulteriori informazioni, consultare l'argomento "Specifica parametri" a pagina 51.
	Aggiunge un nuovo campo	È abilitato solo quando è selezionata una singola riga nella tabella Campi simulati. Aggiunge una nuova riga vuota in fondo alla tabella Campi simulati.

Tabella 5. Icone nella vista Campi simulati (Continua)

Icona	Suggerimento	Descrizione
	Crea più copie	È abilitato solo quando è selezionata una singola riga nella tabella Campi simulati. Consente di aprire la finestra di dialogo secondaria Clona campo. Per ulteriori informazioni, consultare l'argomento "Clona campo" a pagina 49.
	Elimina campo selezionato	Elimina la riga selezionata dalla tabella Campi simulati.
	Sposta in alto	È abilitato solo se la riga selezionata non è già la prima riga della tabella Campi simulati. Sposta la riga selezionata all'inizio della tabella Campi simulati. Questa azione influisce sull'ordine dei campi nei dati simulati.
	Sposta su	È abilitato solo se la riga selezionata non è la prima riga della tabella Campi simulati. Sposta la riga selezionata di una posizione verso l'alto nella tabella Campi simulati. Questa azione influisce sull'ordine dei campi nei dati simulati.
	Sposta giù	È abilitato solo se la riga selezionata non è l'ultima riga della tabella Campi simulati. Sposta la riga selezionata di una posizione verso il basso nella tabella Campi simulati. Questa azione influisce sull'ordine dei campi nei dati simulati.
	Sposta in basso	È abilitato solo se la riga selezionata non è già l'ultima riga della tabella Campi simulati. Sposta la riga selezionata alla fine della tabella Campi simulati. Questa azione influisce sull'ordine dei campi nei dati simulati.

Non cancellare Min e Max durante il riadattamento. Se viene selezionato, i valori minimo e massimo non vengono sovrascritti quando le distribuzioni vengono aggiornate eseguendo un nodo Adattamento simulazione connesso.

Vista Correlazioni

I campi di input per i campi predittivi sono spesso noti per essere correlati, ad esempio, altezza e peso. Le correlazioni tra i campi che verranno simulati devono essere prese in considerazione per poter assicurare che i valori simulati conservino tali correlazioni.

Se il nodo Genera simulazione è stato generato o aggiornato da un nodo Adattamento simulazione utilizzando dati cronologici, nella vista Correlazioni è possibile visualizzare e modificare le correlazioni calcolate tra coppie di campi. Se non si dispone di dati cronologici, è possibile specificare le correlazioni manualmente in base alla propria conoscenza relativa alla modalità di correlazione dei campi.

Nota: Prima di generare qualsiasi dato, la matrice di correlazione viene controllata automaticamente per verificare se è semi definita positiva e può, pertanto, essere invertita. Una matrice può essere invertita se le sue colonne sono indipendenti in modo lineare. Se la matrice di correlazione non può essere invertita, verrà corretta automaticamente per renderla invertibile.

È possibile scegliere di visualizzare le correlazioni in formato matrice o elenco.

Matrice correlazioni. Visualizza le correlazioni tra le coppie di campi in una matrice. I nomi dei campi sono elencati, in ordine alfabetico, a sinistra e in alto lungo la matrice. Solo le celle sotto la diagonale possono essere modificate; è necessario immettere un valore compreso tra -1.000 e 1.000, inclusi. La cella sopra la diagonale viene aggiornata quando il focus viene allontanato dalla relativa cella di mirroring sotto la diagonale; entrambe le celle visualizzeranno quindi lo stesso valore. Le celle della diagonale sono sempre disabilitate e hanno sempre una correlazione 1.000. Il valore di default per tutte le altre celle è 0.000. Il valore 0.000 specifica che non vi è correlazione tra la coppia associata di campi. Solo i campi continui e ordinali sono inclusi nella matrice. I campi nominali, di categoria e indicatore, e i campi a cui viene assegnata la distribuzione fissa non vengono visualizzati nella tabella.

Elenco correlazioni. Visualizza le correlazioni tra le coppie di campi in una tabella. Ogni riga della tabella mostra la correlazione tra una coppia di campi. Non è possibile aggiungere o eliminare righe. Le colonne con le intestazioni Campo 1 e Campo 2 contengono i nomi dei campi, che non possono essere modificati. La colonna Correlazione contiene le correlazioni, che possono essere modificate; è necessario immettere un valore compreso tra -1.000 e 1.000, inclusi. Il valore di default per tutte le celle è 0.000. Solo i campi continui e ordinali sono inclusi nell'elenco. I campi nominali, di categoria e indicatore, e i campi a cui viene assegnata la distribuzione fissa non vengono visualizzati nell'elenco.

Reimposta correlazioni. Consente di aprire la finestra di dialogo Reimposta correlazioni. Se sono disponibili dati cronologici, è possibile scegliere una di tre opzioni:

- **Adattato.** Sostituisce le correlazioni correnti con quelle calcolate utilizzando i dati cronologici.
- **Zeri.** Sostituisce le correlazioni con gli zeri.
- **Annulla.** Chiude la finestra di dialogo. Le correlazioni restano invariate.

Se non sono disponibili dati cronologici, ma sono state apportate modifiche alle correlazioni, è possibile scegliere di sostituire le correlazioni correnti con gli zeri, oppure annullare l'operazione.

Mostra come. Selezionare **Tabella** per visualizzare le correlazioni come matrice. Selezionare **Elenco** per visualizzare le correlazioni come elenco.

Non ricalcolare le correlazioni durante il riadattamento. Selezionare questa opzione se si desidera specificare manualmente le correlazioni ed impedire che vengano sovrascritte durante l'adattamento automatico delle distribuzioni tramite un nodo Adattamento simulazione e i dati cronologici.

Utilizza tabella di contingenza a più vie adattata per gli input con una distribuzione categoriale. Per default, tutti i campi con una distribuzione categoriale sono inclusi in una tabella di contingenza (o una tabella di contingenza a più vie, in base al numero di campi con una distribuzione categoriale). La tabella di contingenza viene costruita, come le correlazioni, quando viene eseguito un nodo Adattamento simulazione. La tabella di contingenza non può essere visualizzata. Quando questa opzione è selezionata, i campi con una distribuzione categoriale vengono simulati utilizzando le percentuali effettive della tabella di contingenza. Ossia, qualsiasi associazione tra campi nominali viene ricreata nei nuovi dati simulati. Quando questa opzione viene deselezionata, i campi con le distribuzioni categoriali vengono simulati utilizzando le percentuali previste della tabella di contingenza. Se si modifica un campo, il campo viene rimosso dalla tabella di contingenza.

Vista Opzioni avanzate

Numero di casi da simulare. Visualizza le opzioni per specificare il numero di casi da simulare e come verranno denominate le eventuali iterazioni.

- **Numero massimo di casi.** Specifica il numero massimo di casi dei dati simulati e i valori di destinazione associati da generare.
- **Iterazioni.** Questo numero viene calcolato automaticamente e non può essere modificato. Un'iterazione viene creata automaticamente ogni volta che un parametro di distribuzione ha più valori specificati.
- **Righe totali.** È abilitato solo quando il numero di iterazioni è superiore a 1. Il numero viene calcolato automaticamente, utilizzando l'equazione visualizzata, e non può essere modificato.
- **Crea campo iterazione.** È abilitato solo quando il numero di iterazioni è superiore a 1. Quando è selezionato, il campo **Nome** è abilitato. Per ulteriori informazioni, consultare l'argomento "Iterazioni" a pagina 52.
- **Nome.** È abilitato solo quando la casella di controllo **Crea campo iterazione** è selezionata, e il numero di iterazioni è superiore a 1. Modificare il nome del campo iterazione immettendo i dati in questo campo di testo. Per ulteriori informazioni, consultare l'argomento "Iterazioni" a pagina 52.

Seme random. L'impostazione di un seme random consente di replicare la simulazione.

- **Replica risultati.** Quando è selezionato, il pulsante **Genera** e il campo **Seme random** sono abilitati.
- **Seme random.** È abilitato solo quando la casella di controllo **Replica risultati** è selezionata. In questo campo è possibile specificare un numero intero da utilizzare come seme random. Il valore di default è 629111597.
- **Genera.** È abilitato solo quando la casella di controllo **Replica risultati** è selezionata. Crea un numero intero pseudo-casuale compreso tra 1 e 999999999, inclusi, nel campo **Seme random**.

Clona campo

Nella finestra di dialogo Clona campo è possibile specificare quante copie creare del campo selezionato e come viene denominata ciascuna copia. È utile disporre di più copie dei campi quando si esaminano gli effetti composti, ad esempio i tassi di interesse o di crescita in un numero di periodi di tempo successivi.

La barra del titolo della finestra di dialogo contiene il nome del campo selezionato.

Numero di copie da effettuare. Contiene il numero di copie del campo da creare. Fare clic sulle frecce per selezionare il numero di copie da creare. Il numero minimo di copie è 1 e il numero massimo è 512. Il numero di copie inizialmente è impostato su 10.

Caratteri suffisso della copia. Contiene i caratteri che vengono aggiunti alla fine del nome del campo di ciascuna copia. Questi caratteri separano il nome del campo dal numero di copia. I caratteri suffissi possono essere modificati immettendo i dati in questo campo. Questo campo può essere lasciato vuoto; in questo caso non vi saranno caratteri tra il nome del campo e il numero della copia. Il carattere di default è un carattere di sottolineatura.

Numero copia iniziale. Contiene il numero suffisso per la prima copia. Fare clic sulle frecce per selezionare il numero della copia iniziale. Il numero minimo della copia iniziale è 1 e il numero massimo è 1000. Il numero della copia iniziale di default è 1.

Incremento numero copia. Contiene l'incremento per i numeri suffisso. Fare clic sulle frecce per selezionare l'incremento. L'incremento minimo è 1 e quello massimo 255. L'incremento inizialmente è impostato su 1.

Campi. Contiene un'anteprima dei nomi campo per le copie, che viene aggiornata se uno dei campi della finestra di dialogo Clona campo viene modificato. Questo testo viene generato automaticamente e non può essere modificato.

OK. Genera tutte le copie secondo quanto specificato nella finestra di dialogo. Le copie vengono aggiunte alla tabella Campi simulati nella finestra di dialogo del nodo Genera simulazione, direttamente sotto la riga che contiene il campo copiato.

Annulla. Chiude la finestra di dialogo. Qualsiasi modifica effettuata viene cancellata.

Dettagli adattamento

La finestra di dialogo Dettagli adattamento è disponibile solo se il nodo Genera simulazione è stato creato o aggiornato eseguendo un nodo Adattamento simulazione. Visualizza i risultati dell'adattamento automatico della distribuzione per il campo selezionato. Le distribuzioni sono ordinate per bontà di adattamento, per cui la distribuzione con adattamento migliore viene elencata per prima. In questa finestra di dialogo è possibile eseguire le seguenti attività:

- Esaminare le distribuzioni adattate ai dati cronologici.
- Selezionare una delle distribuzioni adattate.

Campo. Contiene il nome del campo selezionato. Questo testo non può essere modificato.

Tratta come (misura). Visualizza il tipo di misurazione del campo selezionato. Viene ricavato dalla tabella Campi simulati nella finestra di dialogo del nodo Genera simulazione. Il tipo di misurazione può essere modificato facendo clic sulla freccia e selezionando una misurazione dall'elenco a discesa. Sono disponibili tre opzioni: **Continuo**, **Nominale** e **Ordinale**.

Distribuzioni. La tabella Distribuzioni mostra le distribuzioni appropriate al tipo di misurazione. Le distribuzioni adattate ai dati cronologici sono ordinate in base alla bontà dell'adattamento, dal migliore al peggiore. La bontà dell'adattamento viene determinata dalla statistica di adattamento scelta nel nodo Adattamento simulazione. Le distribuzioni che non sono state adattate ai dati cronologici sono elencate nella tabella in ordine alfabetico sotto le distribuzioni adattate.

La tabella Distribuzione contiene le seguenti colonne:

- **Utilizza.** Il pulsante di opzione selezionato indica quale distribuzione è attualmente scelta per il campo. È possibile sovrascrivere la distribuzione con adattamento migliore selezionando il pulsante di opzione della distribuzione desiderata nella colonna Utilizza. La selezione di un pulsante di opzione nella colonna Utilizza visualizza inoltre un grafico della distribuzione sovrapposta all'istogramma (o grafico a barre) dei dati cronologici per il campo selezionato. È possibile selezionare una sola distribuzione alla volta.
- **Distribuzione.** Contiene il nome della distribuzione. Questa colonna non può essere modificata.
- **Statistiche adattamento.** Contiene le statistiche di adattamento calcolate per la distribuzione. Questa colonna non può essere modificata. Il contenuto della cella dipende dal tipo di misurazione del campo:
 - **Continuo.** Contiene i risultati dei test Anderson-Darling e Kolmogorov-Smirnoff. Inoltre, vengono visualizzati i valori p associati ai test. La statistica di adattamento scelta come criterio di bontà dell'adattamento nel nodo Adattamento simulazione viene visualizzata per prima e viene utilizzata per ordinare le distribuzioni. Le statistiche Anderson-Darling vengono visualizzate come $A=aval$ $P=pval$. Le statistiche Kolmogorov-Smirnoff vengono visualizzate come $K=kval$ $P=pval$. Se una statistica non può essere calcolata, viene visualizzato un punto al posto di un numero.
 - **Nominale e ordinale.** Contiene i risultati del test chi-quadrato. Viene visualizzato inoltre il valore p associato al test. Le statistiche vengono visualizzate come $Chi-Sq=val$ $P=pval$. Se la distribuzione non è stata adattata, viene visualizzato Non adattato. Se la distribuzione non può essere matematicamente adatta, viene visualizzato Non può essere adatta.

Nota: la cella è sempre vuota per la distribuzione Empirica.

- **Parametri.** Contiene i parametri di distribuzione associati ad ogni distribuzione adattata. I parametri vengono visualizzati come $nome_parametro = valore_parametro$, con i parametri separati da un singolo spazio. Per la distribuzione di categoria, i nomi dei parametri sono le categorie e i valori dei parametri sono le probabilità associate. Se la distribuzione non è stata adattata ai dati cronologici, la cella è vuota. Questa colonna non può essere modificata.

Miniatura istogramma. Visualizza un grafico della distribuzione selezionata sovrapposto ad un istogramma dei dati cronologici del campo selezionato.

Miniatura distribuzione. Visualizza una spiegazione ed un'illustrazione della distribuzione selezionata.

OK. Chiude la finestra di dialogo ed aggiorna i valori delle colonne Misurazione, Distribuzione, Parametri e Min, Max della tabella Campi simulati per il campo selezionato con le informazioni della distribuzione selezionata. Viene aggiornata anche l'icona nella colonna Stato per riflettere se la distribuzione selezionata è quella con l'adattamento migliore ai dati.

Annulla. Chiude la finestra di dialogo. Qualsiasi modifica effettuata viene cancellata.

Specifica parametri

Nella finestra di dialogo Specifica parametri è possibile specificare manualmente i valori dei parametri per la distribuzione del campo selezionato. È inoltre possibile scegliere una distribuzione diversa per il campo selezionato.

La finestra di dialogo Specifica parametri può essere aperta in tre modi:

- Fare doppio clic su un nome campo nella tabella Campi simulati nella finestra di dialogo del nodo Genera simulazione.
- Fare clic sulla colonna Parametri o Min, Max della tabella Campi simulati e scegliere **Specifica** dall'elenco.
- Nella tabella Campi simulati, selezionare una riga e fare clic sull'icona **Modifica parametri di distribuzione**.

Campo. Contiene il nome del campo selezionato. Questo testo non può essere modificato.

Distribuzione. Contiene la distribuzione del campo selezionato. Viene ricavata dalla tabella Campi simulati. La distribuzione può essere modificata facendo clic sulla freccia e selezionando una distribuzione dall'elenco a discesa. Le distribuzioni disponibili dipendono dal tipo di archiviazione del campo selezionato.

Lati. Questa opzione è disponibile solo quando è selezionata la distribuzione cubo nel campo **Distribuzione**. Fare clic sulle frecce per selezionare il numero di lati, o categorie, in cui suddividere il campo. Il numero minimo di lati è due e il numero massimo è 20. Il numero di lati inizialmente è impostato su 6.

Parametri di distribuzione. La tabella Parametri di distribuzione contiene una riga per ogni parametro della distribuzione scelta. La tabella contiene due colonne:

- **Parametro.** Contiene i nomi dei parametri. Questa colonna non può essere modificata.
- **Valori.** Contiene i valori dei parametri. Se il nodo Genera simulazione è stato creato o aggiornato da un nodo Adattamento simulazione, le celle in questa colonna contengono i valori di parametro determinati dall'adattamento della distribuzione ai dati cronologici. Se il nodo Genera simulazione è stato aggiunto all'area di disegno dalla tavolozza dei nodi Origine, le celle in questa colonna sono vuote. I valori possono essere modificati immettendo i dati nelle celle. Per ulteriori informazioni sui parametri necessari ad ogni distribuzione e sui valori accettabili dei parametri, consultare l'argomento "Distribuzioni" a pagina 53.

I valori multipli per un parametro devono essere separati da virgole. La specifica di più valori per un parametro definirà più iterazioni della simulazione. È possibile specificare più valori per un solo parametro.

Nota: per i campi con tipi di archiviazione data/ora, è necessario specificare i parametri di distribuzione come interi. Ad esempio, per specificare la data media 1 gennaio 1970, utilizzare il numero intero 0.

Nota: Quando viene selezionata la distribuzione cubo, la tabella Parametri di distribuzione è leggermente diversa. La tabella contiene una riga per ogni lato (o categoria). La tabella contiene una colonna Valore e una colonna Probabilità. La colonna valore contiene un'etichetta per ogni categoria. I valori di default per

le etichette sono gli interi 1-N, dove N è il numero di lati. Le etichette possono essere modificate immettendo i dati nelle celle. È possibile immettere qualsiasi valore nelle celle. Se si desidera utilizzare un valore che non è un numero, il tipo di archiviazione del campo di dati deve essere cambiato in stringa se non è già impostato sulla stringa. La colonna Probabilità contiene la probabilità di ciascuna categoria. Le probabilità non possono essere modificate e vengono calcolate come 1/N.

Anteprima. Mostra un grafico campione della distribuzione, basato sui parametri specificati. Se sono stati specificati due o più valori per un unico parametro, vengono visualizzati i grafici campione di ciascun valore del parametro. Se sono disponibili dati cronologici per il campo selezionato, il grafico della distribuzione viene sovrapposto ad un istogramma dei dati cronologici.

Impostazioni facoltative. Utilizzare queste opzioni per specificare un valore minimo, un valore massimo o entrambi per i dati simulati. I dati simulati inferiori al valore minimo e superiori al valore massimo verranno rifiutati, anche se sarebbero validi per la distribuzione specificata.

- **Specifica minimo.** Selezionare questa opzione per abilitare il campo **Rifiuta valori inferiori**. La casella di controllo è disabilitata se è selezionata la distribuzione Empirica.
- **Rifiuta valori inferiori.** Abilitato solo se è selezionato **Specifica minimo**. Immettere un valore minimo per i dati simulati. Qualsiasi valore simulato inferiore a questo valore verrà rifiutato.
- **Specifica massimo.** Selezionare questa opzione per abilitare il campo **Rifiuta valori superiori**. La casella di controllo è disabilitata se è selezionata la distribuzione Empirica.
- **Rifiuta valori superiori.** Abilitato solo se è selezionato **Specifica massimo**. Immettere un valore massimo per i dati simulati. Qualsiasi valore simulato superiore a questo valore verrà rifiutato.

OK. Chiude la finestra di dialogo ed aggiorna i valori delle colonne Distribuzione, Parametri e Min, Max della tabella Campi simulati per il campo selezionato. Anche l'icona nella colonna Stato viene aggiornata per riflettere la distribuzione selezionata.

Annulla. Chiude la finestra di dialogo. Qualsiasi modifica effettuata viene cancellata.

Iterazioni

Se è stato specificato più di un valore per un campo fisso o un parametro di distribuzione, viene generato un insieme indipendente di casi simulati, di fatto una simulazione separata, per ogni valore specificato. Questo consente di esaminare l'effetto della variazione del campo o del parametro. Ogni insieme di casi simulati viene definito un'*iterazione*. Nei dati simulati, le iterazioni vengono impilate.

Se la casella di controllo **Crea campo di iterazione** nella vista Opzioni avanzate della finestra di dialogo del nodo Genera simulazione è selezionata, viene aggiunto un campo di iterazione ai dati simulati come campo nominale con archiviazione numerica. Il nome di questo campo può essere modificato immettendo i dati nel campo **Nome** nella vista Opzioni avanzate. Questo campo contiene un'etichetta che indica a quale iterazione appartiene ciascun caso simulato. Il formato delle etichette dipende dal tipo di iterazione:

- **Iterazione di un campo fisso.** L'etichetta è il nome del campo, seguito da un segno di uguale, seguito dal valore del campo per l'iterazione, ossia
nome_campo = valore_campo
- **Iterazione di un parametro di distribuzione.** L'etichetta è il nome del campo, seguito dai due punti, seguito dal nome del parametro iterato, seguito da un segno di uguale, seguito dal valore del parametro per quella iterazione, ossia
nome_campo:nome_parametro = valore_parametro
- **Iterazione di un parametro di distribuzione per una distribuzione di categoria o di intervallo.** L'etichetta è il nome del campo, seguito dai due punti, seguito da "Iteration", seguito dal numero di iterazione, ossia
nome_campo: Iteration numero_iterazione

Distribuzioni

È possibile specificare manualmente la distribuzione di probabilità per qualsiasi campo aprendola finestra di dialogo Specifica parametri per quel campo, selezionando la distribuzione desiderata dall'elenco **Distribuzione** ed immettendo i parametri di distribuzione nella tabella **Parametri di distribuzione**. Di seguito sono riportate alcune note relative a distribuzioni particolari.

- **Categoriali.** La distribuzione di categoria descrive un campo di input che ha un numero fisso di valori numerici, definiti categorie. A ogni categoria è associata una probabilità, in modo che la somma delle probabilità in tutte le categorie sia uguale a uno.

Nota: Se si specificano le probabilità per categorie la cui somma non è 1, si riceverà un avviso.

- **Binomiale negativa - Errori.** Descrive la distribuzione del numero di errori in una sequenza di prove prima che venga osservato un numero specificato di esiti positivi. Il parametro *Soglia* è il numero specificato di esiti positivi e il parametro *Probabilità* è la probabilità di esito positivo in una determinata prova.
- **Binomiale negativa - Prove.** Descrive la distribuzione del numero di prove necessarie prima che venga osservato un numero specificato di esiti positivi. Il parametro *Soglia* è il numero specificato di esiti positivi e il parametro *Probabilità* è la probabilità di esito positivo in una determinata prova.
- **Intervallo.** Questa distribuzione è composta da un insieme di intervalli, a ogni intervallo è assegnata una probabilità in modo che la somma delle probabilità in tutti gli intervalli sia uguale a 1. I valori in un determinato intervallo vengono tracciati da una distribuzione uniforme definita in tale intervallo. Gli intervalli vengono specificati immettendo un valore minimo, un valore massimo e una probabilità associata.

Ad esempio, si ritiene che il costo di un materiale grezzo abbia il 40% di possibilità di rientrare nell'intervallo di 10€ - 15€ per unità e il 60% di possibilità che rientri nell'intervallo di 15€ - 20€ per unità. L'utente modellerà il costo con una distribuzione Intervallo composta dai due intervalli [10 - 15] e [15 - 20], impostando la probabilità associata al primo intervallo su 0.4 e la probabilità associata al secondo intervallo su 0.6. Gli intervalli non devono essere contigui e possono anche sovrapporsi. Ad esempio, è possibile specificare gli intervalli 10€ - 15€ e 20€ - 25€ o 10€ - 15€ e 13€ - 16€.

- **Weibull.** Il parametro *Posizione* è un parametro di posizione facoltativo, che specifica dove si trova l'origine della distribuzione.

La tabella riportata di seguito mostra le distribuzioni disponibili per l'adattamento delle distribuzioni personalizzato ed i valori accettabili per i parametri. Alcune di queste distribuzioni sono disponibili per adattamento personalizzato per particolari tipi di archiviazione, anche se non vengono adattate automaticamente a tali tipi di archiviazione dal nodo Adattamento simulazione.

Tabella 6. Distribuzioni disponibili per l'adattamento personalizzato

Distribuzione	Tipo di archiviazione supportato per l'adattamento personalizzato	Parametri	Limiti del parametro	Note
Bernoulli	Intero, reale, dataora	Probabilità	$0 \leq \text{Probabilità} \leq 1$	
Beta	Intero, reale, dataora	Forma 1 Forma 2 Minimo Massimo	≥ 0 ≥ 0 $< \text{Massimo}$ $> \text{Minimo}$	Minimo e massimo sono facoltativi.
Binomiale	Intero, reale, dataora	Numero di prove (n) Probabilità Minimo Massimo	> 0 , intero $0 \leq \text{Probabilità} \leq 1$ $< \text{Massimo}$ $> \text{Minimo}$	Il numero di prove deve essere un intero. Minimo e massimo sono facoltativi.

Tabella 6. Distribuzioni disponibili per l'adattamento personalizzato (Continua)

Distribuzione	Tipo di archiviazione supportato per l'adattamento personalizzato	Parametri	Limiti del parametro	Note
Categoriale	Intero, reale, dataora, stringa	Nome categoria (o etichetta)	$0 \leq \text{Valore} \leq 1$	Il valore è la probabilità della categoria. La somma dei valori deve essere 1, altrimenti viene generato un avviso.
Cubo	Intero, stringa	Lati	$2 \leq \text{Lati} \leq 20$	La probabilità di ogni categoria (lato) è calcolata come $1/N$, dove N è il numero di lati. Le probabilità non possono essere modificate.
Empirica	Intero, reale, dataora			Non è possibile modificare la distribuzione empirica o selezionarla come tipo. La distribuzione empirica è disponibile solo quando sono presenti dati cronologici.
Esponenziale	Intero, reale, dataora	Scala Minimo Massimo	> 0 $< \text{Massimo}$ $> \text{Minimo}$	Minimo e massimo sono facoltativi.
Fissi	Intero, reale, dataora, stringa	Valore		Non è possibile specificare la distribuzione Fissa per ogni campo. Se si desidera che ogni campo nei dati generati sia fisso, è possibile utilizzare un nodo di input utente seguito da un nodo di bilanciamento.
Gamma	Intero, reale, dataora	Forma Scala Minimo Massimo	≥ 0 ≥ 0 $< \text{Massimo}$ $> \text{Minimo}$	Minimo e massimo sono facoltativi.
Lognormale	Intero, reale, dataora	Forma 1 Forma 2 Minimo Massimo	≥ 0 ≥ 0 $< \text{Massimo}$ $> \text{Minimo}$	Minimo e massimo sono facoltativi.
Binomiale negativa - Errori	Intero, reale, dataora	Soglia Probabilità Minimo Massimo	≥ 0 $0 \leq \text{Probabilità} \leq 1$ $< \text{Massimo}$ $> \text{Minimo}$	Minimo e massimo sono facoltativi.

Tabella 6. Distribuzioni disponibili per l'adattamento personalizzato (Continua)

Distribuzione	Tipo di archiviazione supportato per l'adattamento personalizzato	Parametri	Limiti del parametro	Note
Binomiale negativa - Prove	Intero, reale, dataora	Soglia Probabilità Minimo Massimo	≥ 0 $0 \leq \text{Probabilità} \leq 1$ $< \text{Massimo}$ $> \text{Minimo}$	Minimo e massimo sono facoltativi.
Normale	Intero, reale, dataora	Media Deviazione standard Minimo Massimo	≥ 0 > 0 $< \text{Massimo}$ $> \text{Minimo}$	Minimo e massimo sono facoltativi.
Poisson	Intero, reale, dataora	Media Minimo Massimo	≥ 0 $< \text{Massimo}$ $> \text{Minimo}$	Minimo e massimo sono facoltativi.
Intervallo	Intero, reale, dataora	Inizio(X) Fine(X) Probabilità(X)	$0 \leq \text{Valore} \leq 1$	X è l'indice di ciascun bin. La somma dei valori di probabilità deve essere 1.
Triangolare	Intero, reale, dataora	Modalità Minimo Massimo	$\text{Minimo} \leq \text{Valore} \leq \text{Massimo}$ $< \text{Massimo}$ $> \text{Minimo}$	
Uniforme	Intero, reale, dataora	Minimo Massimo	$< \text{Massimo}$ $> \text{Minimo}$	
Weibull	Intero, reale, dataora	Tasso Scala Posizione Minimo Massimo	> 0 > 0 ≥ 0 $< \text{Massimo}$ $> \text{Minimo}$	Posizione, massimo e minimo sono facoltativi.

Nodo input utente

Il nodo Input utente consente di ottenere in modo semplice dati sintetici creandoli oppure modificando dati esistenti. È utile, per esempio, quando si desidera creare un insieme di dati di test per la modellazione.

Creazione di dati

Il nodo Input utente è disponibile nella palette Input e può essere aggiunto direttamente all'area del flusso.

1. Fare clic sulla scheda **Input** della palette dei nodi.
2. Fare doppio clic sul nodo Input utente o trascinarlo per aggiungerlo nell'area del flusso.
3. Fare doppio clic sul nodo aggiunto per aprirne la relativa finestra di dialogo e specificare campi e valori.

Nota: i nodi Input utente selezionati dalla palette Input risulteranno completamente vuoti, senza campi né informazioni sui dati. Ciò consente di creare dati sintetici completamente nuovi.

Creazione di dati da una sorgente dati esistente

È possibile anche generare un nodo Input utente da qualsiasi nodo non terminale del flusso:

1. Stabilire il nodo del flusso che si desidera sostituire.
2. Fare clic con il pulsante destro del mouse sul nodo i cui dati si desidera utilizzare per il nodo Input utente e scegliere **Genera nodo Input utente** dal menu di scelta rapida.
3. Verrà visualizzato il nodo Input utente con tutti i processi di downstream collegati. Tale nodo sostituisce il nodo esistente nel punto stabilito all'interno del flusso di dati. Il nodo Input utente creato eredita dai metadati l'intera struttura di dati e tutte le informazioni sui tipi di campi, se disponibili.

Nota: se i dati non sono stati eseguiti in tutti i nodi del flusso, i nodi risultano non completamente istanziati e, pertanto, è possibile che l'archiviazione e i valori dei dati non siano disponibili quando si sostituisce un nodo esistente con un nodo Input utente.

Impostazioni relative al nodo Input utente

La finestra di dialogo di un nodo Input utente contiene diversi strumenti che è possibile utilizzare per immettere valori e definire la struttura dei dati per i dati sintetici. Per un nodo generato, la tabella nella scheda Dati contiene nomi di campi della sorgente dati originale. Per un nodo aggiunto dalla palette Input, questa tabella risulta invece vuota. Tramite le opzioni della tabella è possibile eseguire le attività riportate di seguito:

- Aggiungere nuovi campi utilizzando il pulsante di aggiunta di un nuovo campo a destra della tabella.
- Rinominare campi esistenti.
- Specificare un tipo di archiviazione per ogni campo.
- Specificare valori.
- Modificare l'ordine dei campi visualizzati.

Inserimento di dati

Per ogni campo è possibile specificare valori o inserire valori dall'insieme di dati originale utilizzando il pulsante di selezione dei valori a destra della tabella. Per ulteriori informazioni sulla definizione di valori, vedere le regole descritte di seguito. È anche possibile decidere di lasciare vuoto il campo — ai campi lasciati vuoti vengono assegnati i valori null di sistema (\$null\$).

Per specificare i valori stringa, immetterli nella colonna Valori, separati da spazi:

Fred Ethel Martin

Le stringhe che includono spazi possono essere racchiuse tra virgolette:

"Bill Smith" "Fred Martin" "Jack Jones"

Per i campi numerici, è possibile immettere più valori nello stesso modo (ovvero separati da spazi):

10 12 14 16 18 20

oppure specificare la serie di numeri immettendo i numeri più basso e più alto (10, 20) e quindi il numero che rappresenta il criterio utilizzato per la serie (2). In base a questo metodo, si specificherà:

10,20,2

È possibile combinare questi metodi utilizzandone uno all'interno dell'altro, per esempio:

1 5 7 10,20,2 21 23

In questo modo si otterranno i valori seguenti:

1 5 7 10 12 14 16 18 20 21 23

I valori di data e ora possono essere immessi utilizzando il formato di default corrente selezionato nella finestra di dialogo Proprietà flusso, per esempio:

11:04:00 11:05:00 11:06:00

2007-03-14 2007-03-15 2007-03-16

Per i valori di timestamp, costituiti da una componente di data e di ora, è necessario utilizzare le virgolette:

"2007-03-14 11:04:00" "2007-03-14 11:05:00" "2007-03-14 11:06:00"

Per ulteriori informazioni, vedere i commenti indicati di seguito relativi all'archiviazione di dati.

Genera dati. Consente di specificare come vengono generati i record quando viene eseguito il flusso.

- **Tutte le combinazioni.** Genera record contenenti ogni possibile combinazione di valori dei campi; in questo modo, il valore di ogni campo comparirà in diversi record. Talvolta questo può generare una quantità di dati eccessiva, pertanto può essere opportuno fare seguire a questo nodo un nodo Campione.
- **In ordine.** Genera i record nell'ordine in cui vengono specificati i valori dei campi di dati. Il valore di ogni campo compare in un solo record. Il numero totale dei record è uguale al numero massimo di valori per un singolo campo. Se i campi hanno un numero di valori inferiore a quello massimo vengono inseriti valori non definiti (\$null\$).

Mostra esempio

Ad esempio, le seguenti voci generano i record elencati nei due esempi di tabella riportati di seguito.

- **Età.** 30,60,10
- **Pressione.** BASSO
- **Colesterolo** NORMALE ALTO
- **Cura.** (vuoto)

Tabella 7. Campo Genera dati impostato su Tutte le combinazioni.

Età	Pressione	Colesterolo	Cura
30	BASSO	NORMAL	\$null\$
30	BASSO	ALTO	\$null\$
40	BASSO	NORMAL	\$null\$
40	BASSO	ALTO	\$null\$
50	BASSO	NORMAL	\$null\$
50	BASSO	ALTO	\$null\$
60	BASSO	NORMAL	\$null\$
60	BASSO	ALTO	\$null\$

Tabella 8. Campo Genera dati impostato su In ordine.

Età	Pressione	Colesterolo	Cura
30	BASSO	NORMAL	\$null\$
40	\$null\$	ALTO	\$null\$
50	\$null\$	\$null\$	\$null\$
60	\$null\$	\$null\$	\$null\$

Archiviazione dei dati

L'archiviazione definisce le modalità di archiviazione dei dati in un campo. Per esempio, un campo con valori 1 e 0 è un campo in cui sono archiviati dati di tipo intero. Si distingue dal livello di misurazione, che descrive l'utilizzo dei dati e non influisce sull'archiviazione. Per esempio, è possibile impostare il

livello di misurazione per un campo intero con valori 1 e 0 su *Flag*. In genere, questo indica che 1=*Vero* e 0=*Falso*. Mentre l'archiviazione deve essere stabilita all'origine, il livello di misurazione può essere modificato tramite un nodo Tipo in qualsiasi punto del flusso. Per ulteriori informazioni, consultare l'argomento "Livelli di misurazione" a pagina 112.

I tipi di archiviazione disponibili sono:

- **Stringa.** Utilizzato per i campi che contengono dati non numerici, detti anche dati alfanumerici. Una stringa può includere qualsiasi sequenza di caratteri, quale *mario*, *Classe 2* o *1234*. Si noti che i numeri presenti nelle stringhe non possono essere utilizzati nei calcoli.
- **Intero.** Un campo i cui valori sono numeri interi.
- **Reale.** I valori sono numeri che possono includere decimali (non limitati ai numeri interi). Il formato di visualizzazione viene specificato nella finestra di dialogo Proprietà flusso e può essere ignorato per singoli campi in un nodo Tipo (scheda Formato).
- **Data.** Valori di data specificati in un formato standard quale anno, mese e giorno (per esempio, 26.09.07). Il formato specifico è indicato nella finestra di dialogo Proprietà flusso.
- **Ora.** Tempo misurato come durata. Per esempio, se una chiamata di assistenza è durata 1 ora, 26 minuti e 38 secondi, potrà essere rappresentata come 01:26:38, in base al formato dell'ora corrente specificato nella finestra di dialogo Proprietà stream.
- **Timestamp.** Valori che includono sia la data che l'ora, ad esempio 2007-09-26 09:04:00, che dipendono dai formati di data ed ora correnti specificati nella finestra di dialogo Proprietà flusso. Notare che potrebbe essere necessario racchiudere i valori timestamp tra doppi apici per garantire che siano interpretati come valore singolo piuttosto che come valori di data ed ora separati. Ad esempio quando si immettono valori in un nodo Input utente.

Conversioni della classe di archiviazione. È possibile convertire la classe di archiviazione di un campo utilizzando una vasta gamma di funzioni di conversione, per esempio *to_string* e *to_integer*, in un nodo Riempimento. Per ulteriori informazioni, consultare l'argomento "Conversione della classe di archiviazione mediante il nodo Riempimento" a pagina 133. Notare che le funzioni di conversione (e tutte le altre funzioni che richiedono un tipo specifico di input, come un valore data o ora) dipendono dai formati correnti specificati nella finestra di dialogo Proprietà flusso. Per esempio, se si desidera convertire un campo stringa con valori *Gen 2003*, *Feb 2003* e così via in campo di archiviazione della data, selezionare **MES AAAA** come formato di data di default per il flusso. Le funzioni di conversione sono disponibili anche dal nodo Ricava per la conversione temporanea durante un calcolo di derivazione. È inoltre possibile utilizzare il nodo Ricava per eseguire altre manipolazioni, quali ricodifiche di campi stringa con valori categoriali. Per ulteriori informazioni, consultare l'argomento "Ricodifica di valori con il nodo Ricava" a pagina 132.

Lettura di dati misti. Quando si leggono campi con archiviazione numerica (numeri interi, reali, ora, timestamp o data), gli eventuali valori non numerici vengono impostati come valori null o valori mancanti di sistema poiché, a differenza di altre applicazioni, IBM SPSS Modeler non consente di utilizzare tipi di archiviazione misti all'interno dello stesso campo. Per evitare questo inconveniente, è necessario modificare opportunamente il tipo di archiviazione nel nodo origine o nell'applicazione esterna per fare in modo che tutti i campi contenenti dati misti siano letti come stringhe.

Nota: i nodi Input utente generati possono già contenere informazioni sull'archiviazione raccolte dal nodo origine, se tale nodo è istanziato. Invece, un nodo non istanziato non conterrà alcuna informazione su archiviazione o tipo di utilizzo.

Regole per la definizione dei valori

Per i campi simbolici, è necessario lasciare uno spazio tra i valori, per esempio:

ALTO MEDIO BASSO

Per i campi numerici, è possibile immettere più valori nello stesso modo (ovvero separati da spazi):

10 12 14 16 18 20

oppure specificare la serie di numeri immettendo i numeri più basso e più alto (10, 20) e quindi il numero che rappresenta il criterio utilizzato per la serie (2). In base a questo metodo, si specificherà:

10,20,2

È possibile combinare questi metodi utilizzandone uno all'interno dell'altro, per esempio:

1 5 7 10,20,2 21 23

In questo modo si otterranno i valori seguenti:

1 5 7 10 12 14 16 18 20 21 23

Schede comuni a tutti i nodi origine

Le opzioni seguenti possono essere specificate per tutti i nodi origine facendo clic sulla scheda corrispondente:

- **Scheda Dati.** Utilizzata per modificare il tipo di archiviazione di default.
- **Scheda Filtro.** Utilizzata per eliminare o rinominare campi di dati. Questa scheda include le stesse opzioni disponibili nel nodo Filtro. Per ulteriori informazioni, consultare l'argomento "Impostazione delle opzioni di filtro" a pagina 122.
- **Scheda Tipi.** Utilizzata per impostare i livelli di misurazione. Questa scheda include le stesse opzioni disponibili nel nodo Tipo.
- **Scheda Annotazioni.** Utilizzata per tutti i nodi, questa scheda fornisce opzioni per rinominare i nodi, uno strumento personalizzato e la possibilità di archiviare annotazioni estese.

Impostazione dei livelli di misurazione nel nodo origine

È necessario specificare le proprietà dei campi in un nodo origine o in un nodo Tipo distinto. Questa funzione è analoga in entrambi i nodi. Sono disponibili le seguenti proprietà:

- **Campo.** Fare doppio clic sul nome di un campo per specificare valore ed etichette di campo per i dati in IBM SPSS Modeler. Per esempio, qui è possibile visualizzare o modificare i metadati dei campi importati da IBM SPSS Statistics. Analogamente, è possibile creare nuove etichette per campi e loro valori. Le etichette specificate qui vengono visualizzate in IBM SPSS Modeler in base alle selezioni effettuate nella finestra di dialogo Proprietà stream.
- **Misurazione.** Livello di misurazione utilizzato per descrivere le caratteristiche dei dati in un campo specifico. Se tutti i dettagli di un campo sono noti, il campo viene definito **completamente istanziato**. Per ulteriori informazioni, consultare l'argomento "Livelli di misurazione" a pagina 112.

Nota: il livello di misurazione di un campo è diverso dal relativo tipo di archiviazione, che indica se i dati sono archiviati come stringhe, interi, numeri reali, date, ore o timestamp.

- **Valori.** Questa colonna consente di specificare le opzioni di lettura dei valori dei dati dall'insieme di dati. L'opzione **Specificata** permette invece di specificare i livelli di misurazione e i valori in una finestra di dialogo a parte. È anche possibile scegliere di passare i campi senza leggerne i valori. Per ulteriori informazioni, consultare l'argomento "Valori dei dati" a pagina 115.
- **Mancante/i.** Utilizzata per specificare come vengono gestiti i valori mancanti del campo. Per ulteriori informazioni, consultare l'argomento "Definizione dei valori mancanti" a pagina 118.
- **Controllo.** In questa colonna è possibile impostare opzioni per assicurarsi che i valori dei campi siano inclusi nei valori o nell'intervallo specificato. Per ulteriori informazioni, consultare l'argomento "Controllo dei valori del tipo" a pagina 119.

- **Ruolo.** Proprietà utilizzata per indicare ai nodi Modelli se i campi verranno trattati come campi di **Input** (campi predittori) o **Obiettivo** (campi previsti) nel processo di apprendimento automatico. **Entrambi** e **Nessuno** sono altri ruoli possibili, oltre a **Partizione**, che indica un campo utilizzato per eseguire la partizione dei record in campioni separati per addestramento, test e convalida. Il valore **Suddivisione** specifica che i modelli separati verranno costruiti per ciascun valore possibile del campo. Per ulteriori informazioni, consultare l'argomento "Impostazione del ruolo del campo" a pagina 119.

Per ulteriori informazioni, consultare l'argomento "Nodo Tipo" a pagina 111.

Istanziamento nel nodo origine

Per ottenere informazioni sull'archiviazione dei dati e sui valori dei campi sono disponibili due metodi di istanziamento. La **creazione istanza** può essere eseguita a livello del nodo origine quando i dati vengono importati in IBM SPSS Modeler o tramite l'aggiunta di un nodo Tipo nel flusso di dati.

L'istanziamento a livello del nodo origine risulta utile nei casi seguenti:

- Quando l'insieme di dati è di dimensioni ridotte.
- Quando si stabilisce di derivare nuovi campi utilizzando il generatore di espressioni (l'istanziamento rende i valori dei campi disponibili mediante il generatore di espressioni).

In genere, se l'insieme di dati non è di grandi dimensioni e non si prevede di aggiungere campi nel flusso in seguito, l'istanziamento a livello del nodo origine risulta il metodo più vantaggioso.

Filtro di campi dal nodo origine

La scheda Filtro nella finestra di dialogo di un nodo origine consente di escludere campi dalle operazioni a valle in base a un'analisi iniziale dei dati. Questo risulta utile, per esempio, se nei dati sono presenti campi duplicati o se si conoscono i dati in modo tanto approfondito da poter escludere i campi irrilevanti. In alternativa, è possibile aggiungere un nodo Filtro separato più avanti nel flusso. Questa funzione è analoga in entrambi i casi. Per ulteriori informazioni, consultare l'argomento "Impostazione delle opzioni di filtro" a pagina 122.

Capitolo 3. Nodi Operazioni su record

Panoramica sui nodi Operazioni su record

I nodi Operazioni su record vengono utilizzati per apportare modifiche ai dati a livello di record. Queste operazioni risultano importanti durante le fasi **Data Understanding** e **Data Preparation** del processo di data mining, poiché consentono di adattare i dati alle specifiche esigenze di business.

In base ai risultati dell'esplorazione dati eseguita utilizzando il nodo Esplora (palette Output), per esempio, è possibile decidere se si desidera unire i record relativi agli acquisti effettuati da un cliente negli ultimi tre mesi. Utilizzando un nodo Unione sarà possibile unire i record in base ai valori di un campo chiave, quale *ID cliente*. Oppure, nel caso in cui non fosse più possibile gestire le informazioni di un database contenente oltre un milione di record relativi ai risultati di un sito Web, utilizzando un nodo Campione è possibile selezionare un sottoinsieme di dati da utilizzare per la modellazione.

La palette Operazioni su record contiene i seguenti nodi:



Il nodo Seleziona consente di selezionare o scartare un sottoinsieme di record dal flusso dei dati basato su una condizione specifica. Per esempio, è possibile selezionare i record relativi a una determinata area vendite.



Il nodo Campione seleziona un sottoinsieme di record. Sono supportati vari tipi di campioni, inclusi campioni stratificati, raggruppati e non casuali (strutturati). Il campionamento può essere utile per migliorare le prestazioni e per selezionare gruppi di record correlati o transazioni per un'analisi.



Il nodo bilanciamento corregge squilibri in un insieme di dati in modo che soddisfi una determinata condizione. La direttiva di bilanciamento regola la proporzione di record in cui una condizione è vera in base al fattore specificato.



Il nodo Aggregazione sostituisce una sequenza di record di input con record di output aggregati di riepilogo.



Il nodo Aggregazione RFM (Recency, Frequency, Monetary, Passato recente, Frequenza, Monetario) consente di prendere in considerazione i dati cronologici delle transazioni dei clienti, eliminare i dati non utilizzati e combinare tutti i dati delle transazioni rimanenti in un'unica riga che indica quanto tempo è trascorso dall'ultima transazione, il numero di transazioni effettuate e il valore monetario totale delle transazioni.



Il nodo Ordina ordina record in ordine crescente o decrescente in base ai valori di uno o più campi.



Il nodo Unione prende più record di input e crea un singolo record di output contenente tutti o alcuni campi di input. È utile per unire dati da sorgenti diverse, per esempio dati interni sui clienti e dati demografici acquistati.



Il nodo Accodamento concatena insieme di record. Può essere utilizzato per combinare insieme di dati con strutture simili ma dati diversi.



Il nodo Elimina duplicati rimuove record duplicati passando il primo record distinto nel flusso di dati oppure scartando il primo record e passando nel flusso tutti i duplicati.



Il nodo Streaming TS crea modelli di serie temporali e ne calcola il punteggio in un unico passo, senza bisogno di un nodo Intervalli di tempo.

Molti di questi nodi richiedono l'utilizzo di espressioni CLEM. Se si conosce il linguaggio CLEM, è possibile immettere un'espressione nel campo. Tuttavia, tutti i campi per le espressioni presentano un pulsante che apre il Generatore di espressioni CLEM, che aiuta a creare tali espressioni in modo automatico.



Figura 1. Pulsante del Generatore di espressioni

Nodo Selezione

È possibile utilizzare i nodi Selezione per selezionare o scartare un sottoinsieme di record dal flusso di dati in base a una condizione specificata, per esempio `Pressione (pressione sanguigna) = "ALTO"`.

Modalità. Specifica se i record che soddisfano la condizione verranno inclusi nel flusso di dati oppure esclusi.

- **Includi.** Selezionare questa opzione per includere i record che soddisfano la condizione di selezione.
- **Scarta.** Selezionare questa opzione per escludere i record che soddisfano la condizione di selezione.

Condizione. Visualizza la condizione di selezione che verrà utilizzata per la verifica di ogni record. Per specificare questa condizione è necessario utilizzare un'espressione CLEM. Immettere un'espressione nella finestra oppure utilizzare il Generatore di espressioni facendo clic sul pulsante con la calcolatrice (Generatore espressioni) a destra della finestra.

Se si decide di eliminare i record in base a una condizione, per esempio la seguente:

```
(var1='value1' and var2='value2')
```

il nodo Selezione elimina per default anche i record con valori null per tutti i campi di selezione. Per evitare ciò, accodare la seguente condizione a quella originale:

```
and not(@NULL(var1) and @NULL(var2))
```

I nodi Seleziona vengono inoltre utilizzati per selezionare una proporzione di record. In genere, per questa operazione si utilizza un nodo diverso, ovvero il nodo Campione. Se tuttavia la condizione che si desidera specificare risulta più complessa dei parametri disponibili, è possibile creare una condizione personalizzata utilizzando il nodo Seleziona. Per esempio, è possibile creare la condizione seguente:

```
BP = "HIGH" and random(10) <= 4
```

In questo modo, circa il 40% dei record contenenti valori di pressione alta verrà selezionato e passato a valle per ulteriori analisi.

Nodo Campione

È possibile utilizzare i nodi Campione per selezionare un sottoinsieme di record per l'analisi, oppure per specificare una percentuale di record da scartare. Sono supportati vari tipi di campioni, inclusi campioni stratificati, raggruppati e non casuali (strutturati). Il campionamento può essere utilizzato per vari motivi:

- Per migliorare le prestazioni stimando i modelli in un sottoinsieme di dati. I modelli stimati a partire da un campione offrono spesso lo stesso livello di precisione di quelli derivati dall'insieme di dati completo e molti di essi sono anche più precisi; le migliori prestazioni consentono così di sperimentare metodi diversi che altrimenti non verrebbero tentati.
- Per selezionare gruppi di record o di transazioni correlati per l'analisi, per esempio per selezionare tutti gli elementi di un carrello (o panier) per gli acquisti online, oppure tutte le proprietà in una specifica vicinanza.
- Per identificare unità o casi per l'ispezione casuale a fini di controllo qualità, prevenzione delle frodi o sicurezza.

Nota: se si desidera semplicemente ripartire i dati in campioni di addestramento e di test ai fini della convalida, è possibile utilizzare invece un nodo Partizione. Per ulteriori informazioni, consultare l'argomento "Nodo Partizione" a pagina 146.

Tipi di campioni

Campioni raggruppati. Gruppi o cluster di campioni anziché unità singole. Per esempio, si supponga di avere un file di dati con un record per studente. Se si effettua il raggruppamento per scuola e la dimensione del campione è 50%, verrà scelto il 50% delle scuole e saranno presi in considerazione tutti gli studenti di ciascuna scuola selezionata. Gli studenti delle scuole non selezionate saranno rifiutati. In media, ci si aspetta che venga selezionato il 50% degli studenti, ma poiché le scuole hanno dimensioni diverse, la percentuale potrebbe non essere esatta. Analogamente, è possibile raggruppare gli articoli del carrello in base all'ID transazione per essere certi che siano mantenuti tutti gli articoli delle transazioni selezionate. Per un esempio di raggruppamento delle proprietà in base alla provincia, consultare il flusso di esempio *complexsample_property.str*.

Campioni stratificati. Campioni selezionati in modo indipendente all'interno di sottogruppi di popolazione che non si sovrappongono, o strati. Per esempio, è possibile fare in modo che uomini e donne siano campionati in proporzioni uguali, oppure che sia rappresentata ogni regione o gruppo socio-economico all'interno di una popolazione urbana. È anche possibile specificare una dimensione diversa del campione per ogni strato (per esempio, se si pensa che un gruppo sia stato sottorappresentato nei dati originali). Per un esempio di stratificazione delle proprietà in base alla regione, consultare il flusso di esempio *complexsample_property.str*.

Campionamento sistematico o 1-ogni-n. Se è difficile ottenere la selezione casuale, è possibile campionare le unità in modo sistematico (a un intervallo fisso) o sequenziale.

Pesi di campionamento. I pesi di campionamento vengono calcolati automaticamente durante la formulazione di un campione complesso e corrispondono approssimativamente alla "frequenza" con cui ogni unità campionata è rappresentata nei dati originali. Pertanto, la somma dei pesi nel campione stima le dimensioni dei dati originali.

Lista di campionamento

Una lista di campionamento definisce la fonte potenziale di casi da includere in un campione o studio. In alcuni casi può essere possibile identificare ogni singolo membro di una popolazione e includerlo in un campione, per esempio durante il campionamento degli articoli risultanti da una linea di produzione. Più spesso, non sarà possibile accedere a ogni caso possibile. Per esempio, non è possibile essere certi di chi parteciperà a una consultazione elettorale fino a dopo la votazione. In questo caso, è possibile utilizzare il registro elettorale come lista di campionamento, anche se alcune persone registrate non voteranno ed altre potrebbero votare pur non essendo inserite nell'elenco al momento della verifica del registro. Le persone non incluse nella lista di campionamento non potranno essere campionate. Se la lista di campionamento rispecchi fedelmente la natura della popolazione che si sta tentando di valutare è una questione che andrà gestita nei singoli casi della vita reale.

Opzioni del nodo Campione

È possibile scegliere il metodo **Semplice** o **Complesso** in base alle esigenze.

Opzioni di campionamento semplice

Il metodo di campionamento Semplice consente di selezionare una percentuale casuale di record, di selezionare record contigui oppure di selezionare ogni N record.

Modalità. Selezionare l'opzione desiderata per passare (includere) o scartare (escludere) record:

- **Includi campione.** Include i record selezionati nel flusso di dati e scarta tutti gli altri. Per esempio, se si seleziona **Includi campione** come Modalità e si imposta l'opzione **1-ogni-n** su 5, verrà incluso un record ogni cinque, generando un insieme di dati che corrisponde approssimativamente a un quinto della dimensione originale. Questa è la modalità di default durante il campionamento dei dati e l'unica modalità durante l'utilizzo del metodo complesso.
- **Scarta campione.** Esclude i record selezionati e include tutti gli altri. Per esempio, se si seleziona **Scarta campione** e si imposta l'opzione **1-ogni-n** su 5, verrà scartato un record ogni cinque. Questa modalità è disponibile solo con il metodo semplice.

Esempio. Specificare il metodo di campionamento selezionando una delle opzioni seguenti:

- **Primo.** Selezionare questa opzione per utilizzare un campionamento di dati contigui. Per esempio, se la dimensione massima del campione è impostata su 10000, verranno selezionati i primi 10000 record.
- **1-ogni-n.** Selezionare questa opzione per campionare dati passando o scartando un record ogni n record. Per esempio, se N è impostato su 5, verrà selezionato un record su cinque.
- **% casuale.** Selezionare questa opzione per campionare una percentuale casuale dei dati. Per esempio, se si imposta la percentuale su 20, il 20% dei dati verrà passato nel flusso di dati o scartato dal flusso, a seconda della modalità selezionata. Utilizzare il campo disponibile per specificare una percentuale di campionamento. È inoltre possibile specificare un valore seme mediante il controllo **Imposta seme random**.

Utilizza campionamento block level (solo in-database). Questa opzione è attivata solo se si sceglie il campionamento con percentuale casuale quando si esegue il mining nel database su un database Oracle o IBM DB2. In questi casi, il campionamento block level può essere più efficiente.

Dimensione massima campione. Specifica il numero massimo di record da includere nel campione. Questa opzione è ridondante e risulta pertanto disattivata se vengono selezionati **Primi** e **Includi**. Si noti inoltre che, se utilizzata in combinazione con l'opzione **% casuale**, questa impostazione può impedire la selezione di determinati record. Per esempio, se un insieme di dati contiene 10 milioni di record e si

seleziona il 50% dei record dopo avere impostato una dimensione massima campione di 3 milioni di record, verrà selezionato solo il 50% dei primi 6 milioni di record e i restanti 4 milioni non avranno alcuna possibilità di essere selezionati. Per evitare questa limitazione, selezionare il metodo di campionamento **Complesso** e richiedere un campione casuale di tre milioni di record senza specificare alcuna variabile cluster o di stratificazione.

Opzioni di campionamento complesso

Le opzioni di campionamento complesso consentono un controllo più preciso del campione che comprende campioni raggruppati, stratificati e ponderati unitamente ad altre opzioni.

Cluster e stratifica. Consente di specificare campi di raggruppamento, stratificazione e peso di input, se necessario. Per ulteriori informazioni, consultare l'argomento "Impostazioni di raggruppamento e stratificazione" a pagina 66.

Tipo di campionamento.

- **Casuale.** Seleziona casualmente i cluster o i record all'interno di ogni strato.
- **Sistematico.** Seleziona i record a intervalli fissi. Questa opzione funziona come il metodo *1-ogni-n*, tranne che per la posizione del primo record, che cambia in base a un seme random. Il valore di *N* viene determinato automaticamente in base alla dimensione o alla proporzione del campione.

Unità campione. È possibile selezionare proporzioni o conteggi come unità campione di base.

Dimensione campione. La dimensione campione può essere specificata in vari modi:

- **Fisso.** Consente di specificare la dimensione globale del campione, sotto forma di conteggio o proporzione.
- **Personalizzato.** Consente di specificare la dimensione del campione per ogni sottogruppo o strato. Questa opzione è disponibile solo se è stato specificato un campo di stratificazione nella sottofinestra di dialogo Cluster e stratifica.
- **Variabile.** Consente all'utente di scegliere un campo che definisce la dimensione del campione per ogni sottogruppo o strato. Questo campo deve presentare lo stesso valore per ogni record all'interno di un particolare strato; per esempio, se il campione è stratificato per regione, tutti i record con *regione = Lazio* devono presentare lo stesso valore. Il campo deve essere numerico e i suoi valori devono corrispondere alle unità campione selezionate. Per le proporzioni, i valori devono essere superiori a 0 e inferiori a 1; per i conteggi, il valore minimo è 1.

Campione minimo per strato. Specifica un numero minimo di record (oppure un numero minimo di cluster, se è specificato un campo cluster).

Campione massimo per strato. Specifica un numero massimo di record o cluster. Se si seleziona questa opzione senza specificare un campo cluster o di stratificazione, verrà selezionato un campione casuale o sistematico delle dimensioni specificate.

Imposta seed random. Quando si esegue il campionamento o il partizionamento di record in base a una percentuale casuale, questa opzione consente di duplicare gli stessi risultati in un'altra sessione. Specificando il valore iniziale utilizzato dal generatore di numeri random, è possibile garantire che vengano assegnati gli stessi record a ogni esecuzione del nodo. Immettere il valore seme desiderato, oppure fare clic sul pulsante **Genera** per generare automaticamente un valore casuale. Se questa opzione non è selezionata, verrà generato un campione diverso ogni volta che si esegue il nodo.

Nota: quando si utilizza l'opzione **Imposta seme random** con record letti da un database, potrebbe essere necessario un nodo Ordina prima di eseguire il campionamento, per garantire lo stesso risultato ogni volta che viene eseguito il nodo. Questo si verifica perché il seme random dipende dall'ordine dei record, per il quale non si ha la garanzia che rimanga invariato in un database relazionale. Per ulteriori informazioni, consultare l'argomento "Nodo Ordina" a pagina 73.

Impostazioni di raggruppamento e stratificazione

La finestra di dialogo Cluster e stratifica consente di selezionare i campi cluster, di stratificazione e di peso durante la creazione di un campionamento complesso.

Cluster. Specifica un campo categoriale utilizzato per raggruppare i record. I record vengono campionati in base all'appartenenza ai cluster, con alcuni cluster inclusi e altri no. Tuttavia, se un record qualsiasi di un dato cluster è incluso, tutti i record lo sono. Per esempio, durante l'analisi delle associazioni di prodotti nei carrelli, è possibile raggruppare gli articoli in base all'ID transazione per assicurarsi che siano mantenuti tutti gli articoli delle transazioni selezionate. Invece di campionare i record — eliminando le informazioni relative agli articoli venduti insieme — è possibile campionare le transazioni in modo che tutti i record per le transazioni selezionate vengano conservati.

Stratifica per. Specifica un campo categoriale utilizzato per stratificare i record in modo che i campioni vengano selezionati in modo indipendente all'interno di sottogruppi della popolazione che non si sovrappongono, o strati. Per esempio, se si seleziona un campione del 50% stratificato in base al sesso, verranno considerati due campioni del 50%, uno per gli uomini e uno per le donne. Gli strati possono essere, per esempio, gruppi socioeconomici, categorie professionali, fasce di età o gruppi etnici, consentendo adeguate dimensioni di campionamento per i sottogruppi di interesse. Se nell'insieme di dati vi è un numero di donne tre volte superiore al numero di uomini, questo rapporto verrà mantenuto dal campionamento separato di ogni gruppo. È anche possibile specificare più campi di stratificazione (per esempio, campionamento delle linee di prodotti all'interno delle regioni o viceversa).

Nota: se si effettua la stratificazione in base a un campo con valori mancanti (valori null o mancanti di sistema, stringhe vuote, spazi vuoti e valori vuoti o mancanti definiti dall'utente), non è possibile specificare dimensioni di campionamento personalizzate per gli strati. Se si desidera utilizzare dimensioni di campionamento personalizzate durante la stratificazione in base a un campo con valori mancanti o vuoti, è necessario compilarli a monte.

Utilizza peso di input. Specifica un campo utilizzato per pesare i record prima del campionamento. Per esempio, se il campo peso presenta valori da 1 a 5, i record con peso 5 hanno cinque volte più probabilità di essere selezionati. I valori di questo campo verranno sovrascritti dai pesi di output finali generati dal nodo (vedere il paragrafo successivo).

Nuovo peso di output. Specifica il nome del campo in cui vengono scritti i pesi finali se non viene specificato alcun campo peso di input. Se è specificato un campo peso di input, i suoi valori vengono sostituiti dai pesi finali come indicato sopra e non viene creato alcun campo peso di output separato. I valori del peso di output indicano il numero di record rappresentato da ogni record campionato nei dati originali. La somma dei valori dei pesi fornisce una stima delle dimensioni del campione. Per esempio, se viene considerato un campione casuale del 10%, il peso di output sarà 10 per tutti i record, a indicare che ogni record campionato rappresenta approssimativamente dieci record nei dati originali. In un campione stratificato o ponderato, i valori del peso di output possono variare in funzione della proporzione del campione di ogni strato.

Commenti

- Il campionamento raggruppato è utile se non è possibile ottenere un elenco completo della popolazione che si desidera campionare ma si possono ottenere elenchi completi di determinati gruppi o cluster. Viene usato anche nei casi in cui un campione casuale produrrebbe un elenco di soggetti di test che risulterebbe impossibile contattare. Per esempio, risulterebbe più semplice visitare tutti gli imprenditori agricoli di una regione anziché un elenco selezionato di imprenditori agricoli sparsi in ogni regione del paese.
- È possibile specificare sia campi cluster che di stratificazione in modo da campionare i cluster in modo indipendente all'interno di ogni strato. Per esempio, è possibile campionare i valori delle proprietà stratificati per regione e raggruppati in base alle province di ogni regione. Ciò garantisce la

formulazione di un campione indipendente di province da ogni regione. Alcune province verranno incluse e altre no, ma per ogni provincia inclusa verranno incluse tutte le proprietà all'interno della provincia stessa.

- Per selezionare un campione casuale di unità da ogni cluster, è possibile unire due nodi Campione. Per esempio, è possibile prima campionare le province stratificate per regione come descritto sopra, e quindi collegare un secondo nodo Campione e selezionare *provincia* come campo di stratificazione, per campionare una proporzione di record da ogni provincia.
- Nei casi in cui sia necessaria una combinazione di campi per identificare i cluster in modo univoco, è possibile generare un nuovo campo utilizzando un nodo Ricava. Per esempio, se più negozi utilizzano lo stesso sistema di numerazione delle transazioni, è possibile creare un nuovo campo che concatena gli ID di negozi e transazioni.

Dimensioni dei campioni per gli strati

Durante la formulazione di un campione stratificato, l'opzione di default consente di campionare la stessa proporzione di record o cluster di ogni strato. Se un gruppo supera un altro di un fattore di 3, per esempio, in genere si desidera mantenere lo stesso rapporto nel campione. In caso contrario, tuttavia, è possibile specificare la dimensione del campione in modo separato per ogni strato.

La finestra di dialogo Dimensioni campione per gli strati elenca i singoli valori del campo di stratificazione, consentendo di escludere i valori di default per quel dato strato. Se sono selezionati più campi di stratificazione, viene elencata ogni possibile combinazione di valori, consentendo di specificare, per esempio, la dimensione di ogni gruppo etnico all'interno di ogni provincia, oppure di ogni provincia di ogni regione. Le dimensioni vengono specificate come proporzioni o conteggi, in base all'impostazione corrente nel nodo Campione.

Per specificare le dimensioni dei campioni per gli strati

1. Nel nodo Campione, selezionare **Complesso** e selezionare uno o più campi di stratificazione. Per ulteriori informazioni, consultare l'argomento "Impostazioni di raggruppamento e stratificazione" a pagina 66.
2. Selezionare **Personalizzato** e quindi **Specifica dimensioni**.
3. Nella finestra di dialogo Dimensioni campione per gli strati, selezionare il pulsante **Leggi valori** in basso a sinistra per popolare la visualizzazione. Se necessario, è possibile istanziare i valori in un nodo origine o Tipo a monte. Per ulteriori informazioni, consultare l'argomento "Che cos'è la creazione istanza?" a pagina 114.
4. Fare clic su una riga qualsiasi per ignorare la dimensione di default dello strato.

Note sulle dimensioni dei campioni

Le dimensioni personalizzate dei campioni possono essere utili se strati diversi hanno varianze diverse, per esempio per rendere le dimensioni dei campioni proporzionali alla deviazione standard (se i casi all'interno dello strato sono più vari, è necessario campionarne un numero maggiore per ottenere un campione significativo). Se invece uno strato è di piccole dimensioni, può essere utile utilizzare una proporzione di campionamento superiore per assicurarsi che venga incluso un numero minimo di osservazioni.

Nota: se si effettua la stratificazione in base a un campo con valori mancanti (valori null o mancanti di sistema, stringhe vuote, spazi vuoti e valori vuoti o mancanti definiti dall'utente), non è possibile specificare dimensioni di campionamento personalizzate per gli strati. Se si desidera utilizzare dimensioni di campionamento personalizzate durante la stratificazione in base a un campo con valori mancanti o vuoti, è necessario compilarli a monte.

Nodo bilanciamento

È possibile utilizzare i nodi bilanciamento per correggere squilibri negli insiemi di dati affinché soddisfino i criteri di test specificati. Per esempio, si supponga che un dataset contenga solo due valori, *basso* o *alto*, e che per il 90% dei casi il valore sia *basso* e solo per il 10% sia *alto*. Molte tecniche di modellazione utilizzeranno dati distorti di questo tipo in modo non corretto, infatti tenderanno ad apprendere solo i risultati con valore *basso* e ignorare i risultati con valore *alto*, dal momento che risulta più raro. Se i dati sono bilanciati correttamente con un numero quasi uguale di risultati con valore *basso* e *alto*, sarà più probabile che i modelli riescano a individuare schemi distintivi per i due gruppi. In questo caso, è possibile utilizzare un nodo bilanciamento per creare una direttiva di bilanciamento in grado di ridurre i casi contenenti un risultato con valore *basso*.

Il bilanciamento viene eseguito tramite la duplicazione e il successivo scarto di record in base alle condizioni specificate. I record che non soddisfano alcuna condizione vengono passati nel flusso. Dal momento che questo processo funziona tramite duplicazione e/o scarto di record, la sequenza originale dei dati viene persa nelle operazioni a valle. Pertanto, prima di aggiungere un nodo bilanciamento al flusso di dati è necessario assicurarsi che ogni valore correlato alla sequenza venga derivato.

Nota: i nodi bilanciamento possono essere generati automaticamente da istogrammi e grafici di distribuzione. Per esempio, è possibile bilanciare i dati per mostrare proporzioni uguali in tutte le categorie di un campo categoriale, come evidenziato in un plot di distribuzione.

Esempio. Durante la creazione di un flusso RFM per individuare i clienti recenti che hanno risposto in modo favorevole alle precedenti campagne di marketing, la divisione marketing di un'azienda utilizza un nodo bilanciamento per bilanciare le differenze tra le risposte vere e false nei dati.

Impostazione delle opzioni per il nodo bilanciamento

Direttive di bilanciamento record. Elenca le direttive di bilanciamento correnti. Ogni direttiva include un fattore e una condizione con cui è possibile indicare al software di "aumentare la proporzione di record in base a un fattore specificato se la condizione è vera". Un fattore inferiore a 1,0 indica che la proporzione di record specificati verrà diminuita. Per esempio, se si desidera diminuire il numero di record in cui come trattamento viene utilizzata la cura denominata curaY, è possibile creare una direttiva di bilanciamento con un fattore di 0,7 e una condizione Cura = "curaY". Questa direttiva indica che il numero di record in cui il trattamento utilizzato è curaY verrà ridotto del 70% per tutte le operazioni a valle.

Nota: per la riduzione è possibile specificare fattori di bilanciamento contenenti fino a quattro decimali. I fattori impostati su valori inferiori a 0,0001 determineranno un errore, poiché i risultati non vengono calcolati correttamente.

- **Creare le condizioni** facendo clic sul pulsante a destra del campo testo. In questo modo verrà aggiunta una riga vuota in cui è possibile specificare le nuove condizioni. Per creare un'espressione CLEM per la condizione, fare clic sul pulsante del builder di espressioni.
- **Eliminare le direttive** utilizzando il pulsante con la X rossa.
- **Ordinare le direttive** utilizzando i pulsanti con le frecce.

Bilancia solo dati di addestramento. Se nel flusso è presente un campo partizione, questa opzione bilancia solo i dati nella partizione di addestramento. Questo può essere utile, in particolare, se si generano punteggi di propensione regolata che richiedono una partizione di test o di convalida non bilanciata. Se nel flusso non è presente alcun campo partizione (o se sono specificati più campi partizione), questa opzione viene ignorata e vengono bilanciati tutti i dati.

Nodo Aggregazione

L'aggregazione costituisce un'attività di preparazione dei dati utilizzata spesso per ridurre le dimensioni di un dataset. Prima di procedere all'aggregazione è opportuno eseguire la pulitura dei dati, prestando particolare attenzione ai valori mancanti. Con l'aggregazione potrebbero andare perse informazioni potenzialmente utili sui valori mancanti.

È possibile utilizzare un nodo Aggregazione per sostituire una sequenza di record di input con record di output aggregati di riepilogo. Ad esempio, è possibile che sia disponibile una serie di record delle vendite di input come quelle riportate nella seguente tabella.

Tabella 9. Esempio di input di record delle vendite

Età	Sesso	Regione	Ramo	Vendite
23	M	S	8	4
45	M	S	16	4
37	M	S	8	5
30	M	S	5	7
44	M	N	4	9
25	M	N	2	11
29	F	S	16	6
41	F	N	4	8
23	F	N	6	2
45	F	N	4	5
33	F	N	6	10

È possibile aggregare questi record utilizzando *Sesso* e *Regione* come campi chiave e quindi scegliendo di aggregare *Età* con la modalità **Media** e *Vendite* con la modalità **Somma**. Selezionare **Includi conteggio record nel campo** nella finestra di dialogo Nodo Aggregazione; l'output aggregato sarà simile a quello riportato nella seguente tabella.

Tabella 10. Esempio di record aggregato

Età (media)	Sesso	Regione	Vendite (somma)	Conteggio dei record
35.5	F	N	25	4
29	F	S	6	1
34.5	M	N	20	2
33.75	M	S	20	4

Si apprende, ad esempio, che l'età media delle quattro commesse donne nella regione settentrionale (Nord), sia di 35,5 e che la somma totale delle loro vendite ammonta a 25 unità.

Nota: i campi quali *Ramo* verranno automaticamente scartati se non viene specificata una modalità di aggregazione.

Impostazione delle opzioni per il nodo Aggregazione

Sul nodo Aggregazione si specificano i seguenti elementi.

- Uno o più campi chiave da utilizzare come categorie per l'aggregazione
- Uno o più campi aggregati per i quali calcolare i valori aggregati
- Una o più modalità di aggregazione (tipi di aggregazione) da produrre per ciascun campo aggregato

È inoltre possibile specificare le modalità di aggregazione di default da utilizzare per i nuovi campi aggiunti e utilizzare le espressioni (simili alle formule) per suddividere in categorie l'aggregazione.

Notare che per performance aggiunte, le operazioni di aggregazione potrebbero beneficiare dell'attivazione dell'elaborazione parallela.

Campi chiave. Elenca i campi che è possibile utilizzare come categorie per l'aggregazione. Sia i campi continui (numerici) che campi categoriali possono essere utilizzati come chiavi. Se si selezionano più campi chiave, i valori verranno combinati in modo da ottenere un valore chiave per l'aggregazione dei record. Per ogni campo chiave univoco verrà generato un solo record aggregato. Per esempio, se *Sesso* e *Regione* sono i campi chiave disponibili, verrà generato un record aggregato per ogni combinazione univoca di *M* e *F* con le aree *N* e *S* (ovvero quattro combinazioni univoche). Per aggiungere un campo chiave, utilizzare il pulsante *Selettore di campo* a destra della finestra.

Chiavi contigue. Selezionare questa opzione se si sa che tutti i record con gli stessi valori chiave sono raggruppati insieme nell'input (per esempio, se l'input è ordinato in base ai campi chiave). In questo modo si migliorano le prestazioni.

Il resto della finestra di dialogo è suddivisa in due aree principali - *Aggregati di base* e *Espressioni aggregate*.

Aggregati di base

Campi aggregati. Elenca sia i campi per i quali verranno aggregati i valori, sia le modalità di aggregazione selezionate. Per aggiungere campi a questo elenco, utilizzare il pulsante *Selettore di campo* a destra. Sono disponibili le seguenti modalità di aggregazione.

Nota: alcune modalità non sono applicabili ai campi non numerici (ad esempio, **Somma** per un campo data/ora). Le modalità che non possono essere utilizzate con un campo aggregato selezionato vengono disabilitate.

- **Somma.** Selezionare questa opzione per restituire i valori sommati per ogni combinazione di campi chiave. La somma è il totale di tutti i valori non mancanti di tutti i casi.
- **Media.** Selezionare questa opzione per restituire i valori medi per ogni combinazione di campi chiave. La media è la misurazione della tendenza centrale ed è la media aritmetica (la somma divisa per il numero di casi).
- **Min.** Selezionare questa opzione per restituire i valori minimi per ogni combinazione di campi chiave.
- **Max.** Selezionare questa opzione per restituire i valori massimi per ogni combinazione di campi chiave.
- **SDev.** Selezionare questa opzione per restituire la deviazione standard per ogni combinazione di campi chiave. La deviazione standard è una misura di dispersione intorno alla media ed è uguale alla radice quadrata della misurazione della varianza.
- **Mediana.** Selezionare questa opzione per restituire i valori mediani per ogni combinazione di campi chiave. La mediana è una misura di tendenza centrale che non è sensibile ai valori anomali, a differenza della media che può essere influenzata da valori eccezionalmente bassi o alti. È nota anche come 50° percentile o 2° quartile.
- **Conteggio.** Selezionare questa opzione per restituire il conteggio dei valori non null per ogni combinazione di campi chiave.
- **Varianza.** Selezionare questa opzione per restituire i valori della varianza per ogni combinazione di campi chiave. La varianza è una misura della dispersione dei valori intorno alla media. È calcolata come somma dei quadrati degli scostamenti dalla media, divisa per un valore corrispondente alla somma dei casi meno uno.
- **1° quartile.** Selezionare questa opzione per restituire i valori del primo quartile (venticinquesimo percentile) per ogni combinazione di campi chiave.

- **3° quartile.** Selezionare questa opzione per restituire i valori del terzo quartile (settantacinquesimo percentile) per ogni combinazione di campi chiave.

Nota: durante l'esecuzione di un flusso che contiene un nodo Aggregazione, i valori restituiti per il primo ed il terzo quartile durante il push back del codice SQL ad un database Oracle possono essere diversi da quelli restituiti in modalità nativa.

Modalità di default. Specificare la modalità di aggregazione di default da utilizzare per i campi aggiunti. Se si utilizza spesso la stessa aggregazione, selezionare una o più modalità e utilizzare il pulsante a destra di questo gruppo di opzioni per applicare le modalità selezionate a tutti i campi sopra elencati.

Nuova estensione nome di campo. Selezionare questa opzione per aggiungere un suffisso o un prefisso, per esempio "1" o "nuovo", per duplicare i campi aggregati. Per esempio, un'aggregazione di valori minimi nel campo *Età* per cui è stata selezionata l'opzione di suffisso ed è stato specificato "1" come estensione restituirà il nome di campo *Età_Min_1*. *Nota:* le estensioni di aggregazione quali *_Min* o *_Max* vengono automaticamente aggiunte al nuovo campo, come indicazione del tipo di aggregazione eseguita. Selezionare **Suffisso** o **Prefisso** per indicare il tipo di estensione desiderato.

Includi conteggio record nel campo. Selezionare questa opzione per includere un ulteriore campo in ogni record di risultato. A questo campo verrà assegnato il nome di default *Conteggio_record*. Tale campo indica il numero dei record di input che sono stati aggregati per formare ogni record aggregato. Per assegnare un nome personalizzato a questo campo, digitare il nome desiderato nel campo di modifica.

Nota: i valori di sistema null verranno esclusi dal calcolo degli aggregati, ma inclusi nel conteggio dei record. I valori vuoti, invece, verranno inclusi sia nel calcolo di aggregazione sia nel conteggio dei record. Per escludere i valori vuoti è possibile utilizzare un nodo Riempimento in modo da sostituire i valori vuoti con valori null. È inoltre possibile rimuovere i valori vuoti utilizzando un nodo Seleziona.

Espressioni aggregate

Le espressioni sono simili alle formule che sono state create dai valori, dai nomi del campo, dagli operatori e dalle funzioni. A differenza delle funzioni che lavorano su un singolo record alla volta, le espressioni aggregate funzionano su un gruppo, un insieme o una raccolta di record.

Nota: È possibile creare solo espressioni aggregate se il flusso include una connessione al database (mediante un nodo origine del database).

Nuove espressioni vengono create come campi derivati; per creare un'espressione si utilizzano le funzioni dei *Database Aggregati* che sono disponibili nel Builder di espressioni.

Notare che c'è una connessione tra i **Campi chiave** e qualsiasi espressione aggregata che si crea perchè le espressioni aggregate sono raggruppate per campo chiave.

Le espressioni aggregate valide sono quelle che restituiscono risultati aggregati; un paio di esempi di espressioni di aggregazione valide e le regole che le governano, sono le seguenti:

- È possibile utilizzare le funzioni scalari per associare più funzioni di aggregazione insieme per produrre un singolo risultato di aggregazione. Ad esempio:
`max(C01) - min(C01)`
- Una funzione di aggregazione può operare sul risultato di più funzioni scalari. Ad esempio:
`sum (C01*C01)`

Nodo Aggregazione RFM

Il nodo Aggregazione RFM (Recency, Frequency, Monetary, Passato recente, Frequenza, Monetario) consente di prendere in considerazione i dati cronologici delle transazioni dei clienti, di eliminare i dati non utilizzati e di combinare tutti i dati delle transazioni rimanenti in un'unica riga (utilizzando come chiave l'ID cliente univoco) che elenca la data dell'ultima trattativa (passato recente), il numero di transazioni effettuate (frequenza) e il valore totale delle transazioni (monetario).

Prima di procedere a qualsiasi aggregazione è opportuno eseguire la pulitura dei dati, prestando particolare attenzione ai valori mancanti.

Una volta identificati e trasformati i dati utilizzando il nodo Aggregazione RFM, è possibile utilizzare un nodo Analisi RFM per eseguire ulteriori esplorazioni. Per ulteriori informazioni, consultare l'argomento "Nodo Analisi RFM" a pagina 144.

Si noti che, dopo essere stato elaborato con il nodo Aggregazione RFM, il file di dati non disporrà di alcun valore obiettivo; pertanto, prima di poterlo utilizzare come input per ulteriori analisi predittive con altri nodi Modelli quali C5.0 o CHAID, sarà necessario unirlo ad altri dati di clienti (per esempio, tramite la corrispondenza degli ID cliente). Per ulteriori informazioni, consultare l'argomento "Nodo Unione" a pagina 74.

I nodi Aggregazione RFM e Analisi RFM di IBM SPSS Modeler sono impostati per utilizzare la discretizzazione indipendente, ovvero classificano e discretizzano i dati rispetto a ogni misura di passato recente, frequenza e valore monetario, indipendentemente dai loro valori o dalle altre due misure.

Impostazione delle opzioni per il nodo Aggregazione RFM

La scheda Impostazioni del nodo Aggregazione RFM contiene i campi riportati di seguito.

Calcola passato recente relativo a. Specifica la data a partire dalla quale verrà calcolato il passato recente delle transazioni. Può essere una **Data fissa** inserita dall'utente oppure la **Data odierna** impostata dal sistema. La **Data odierna** è inserita per default e viene aggiornata automaticamente quando viene eseguito il nodo.

ID contigui. Se i dati sono preordinati in modo che tutti i record con lo stesso ID appaiano insieme nel flusso di dati, selezionare questa opzione per accelerare l'elaborazione. Se i dati non sono preordinati (o non si è certi che lo siano), lasciare questa opzione deselezionata e il nodo ordinerà i dati automaticamente.

ID. Selezionare il campo da utilizzare per identificare i clienti e le loro transazioni. Per visualizzare i campi tra i quali è possibile effettuare la selezione, utilizzare il pulsante Selettore di campo a destra.

Data. Selezionare il campo data da utilizzare per calcolare il passato recente. Per visualizzare i campi tra i quali è possibile effettuare la selezione, utilizzare il pulsante Selettore di campo a destra.

Si noti che è richiesto un campo contenente un'archiviazione della data o del timestamp nel formato appropriato da utilizzare come input. Per esempio, se si dispone di un campo stringa con valori come *Gen 2007*, *Feb 2007* e così via, è possibile convertirlo in campo data mediante un nodo Riempimento e la funzione `to_date()`. Per ulteriori informazioni, consultare l'argomento "Conversione della classe di archiviazione mediante il nodo Riempimento" a pagina 133.

Valore. Selezionare il campo da utilizzare per calcolare il valore monetario totale delle transazioni del cliente. Per visualizzare i campi tra i quali è possibile effettuare la selezione, utilizzare il pulsante Selettore di campo a destra. *Nota:* questo valore deve essere numerico.

Nuova estensione nome di campo. Selezionare questa opzione per aggiungere un suffisso o un prefisso, per esempio "12_mesi", ai campi passato recente, frequenza e monetario appena generati. Selezionare **Suffisso** o **Prefisso** per indicare il tipo di estensione desiderato. Questo può essere utile, per esempio, se si esaminano più periodi di tempo.

Scarta record con valore inferiore a. Se necessario, è possibile specificare un valore minimo al di sotto del quale i dettagli della transazione non vengono utilizzati per il calcolo dei totali RFM. Le unità di valore si correlano al campo **Valore** selezionato.

Includi solo transazioni recenti. Se si sta analizzando un database di grandi dimensioni, è possibile specificare che vengano utilizzati solo i record più recenti. È possibile scegliere di utilizzare i dati registrati dopo una certa data o in un periodo di tempo recente:

- **Data transazione successiva a.** Specificare la data della transazione dopo la quale i record verranno inclusi nell'analisi.
- **Transazione entro gli ultimi.** Specificare il numero e il tipo di periodi (giorni, settimane, mesi o anni) prima della data **Calcola passato recente relativo a** dopo la quale i record verranno inclusi nell'analisi.

Salva data della seconda transazione più recente. Selezionare questa casella se si desidera conoscere la data della seconda transazione più recente per ogni cliente. Inoltre, è possibile selezionare anche la casella **Salva data della terza transazione più recente.** Per esempio, questo può essere utile per identificare i clienti che potrebbero aver effettuato più transazioni parecchio tempo fa ma solo una in tempi più recenti.

Nodo Ordina

È possibile utilizzare i nodi Ordina per ordinare i record in ordine crescente o decrescente in base ai valori di uno o più campi. Per esempio, i nodi Ordina vengono spesso utilizzati per visualizzare e selezionare i record con i valori di dati più comuni. In genere, si aggregano innanzitutto i dati mediante il nodo Aggregazione e quindi si utilizza il nodo Ordina per ordinare i dati aggregati secondo un ordine decrescente dei conteggi di record. La visualizzazione di questi risultati in una tabella consentirà di analizzare i dati e di prendere decisioni, per esempio selezionando i record relativi ai 10 migliori clienti.

La scheda Impostazioni del nodo Ordina contiene i campi riportati di seguito.

Ordina per. Vengono visualizzati in una tabella tutti i campi selezionati per essere utilizzati come chiavi di ordinamento. I campi numerici assicurano le migliori prestazioni di ordinamento.

- **Aggiungere campi** a questo elenco utilizzando il pulsante Selettore di campo a destra.
- **Selezionare un ordine** facendo clic sulla freccia **Crescente** o **Decrescente** nella colonna *Ordine* della tabella.
- **Eliminare i campi** utilizzando il pulsante con la X rossa.
- **Ordinare le direttive** utilizzando i pulsanti con le frecce.

Criterio di ordinamento di default. Selezionare **Crescente** o **Decrescente** per utilizzare il relativo criterio di ordinamento come impostazione di default per i nuovi campi aggiunti nella tabella sopra.

Impostazioni di ottimizzazione dell'ordinamento

Se si sta lavorando con dei dati già ordinati in base ad alcuni campi chiave, è possibile specificare quali campi sono già ordinati, consentendo al sistema di ordinare il resto dei dati in modo più efficiente. Questo può accadere, per esempio, se si desidera effettuare l'ordinamento per *Età* (decrescente) e *Cura* (crescente) ma si sa che i dati sono già ordinati per *Età* (decrescente).

I dati sono preordinati. Specifica se i dati sono già ordinati in base a uno o più campi.

Specifica criterio di ordinamento esistente. Specificare i campi già ordinati. Utilizzando la finestra di dialogo Seleziona campi, aggiungere i campi all'elenco. Nella colonna *Ordine*, specificare se ciascun

campo è ordinato in ordine crescente o decrescente. Se si specificano più campi, assicurarsi di elencarli secondo l'ordinamento corretto. Utilizzare le frecce a destra dell'elenco per disporre i campi nell'ordine corretto. In caso di errore nell'indicazione del criterio di ordinamento esistente corretto, al momento dell'esecuzione del flusso verrà visualizzato un errore che indica il numero dei record nei quali l'ordinamento è incoerente rispetto all'impostazione specificata.

Nota: la velocità di ordinamento può beneficiare dell'attivazione dell'elaborazione parallela.

Nodo Unione

Un nodo Unione consente di utilizzare più record di input per creare un singolo record di output contenente alcuni o tutti i campi di input. Questa operazione risulta utile se si desidera unire dati da sorgenti diverse, per esempio dati interni sui clienti e dati demografici acquisiti dall'esterno. È possibile unire i dati nei seguenti modi.

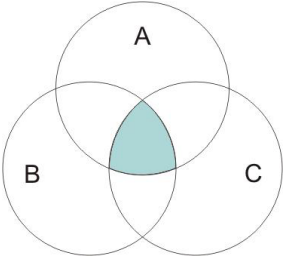
- **Unione per ordine** concatena i record corrispondenti da tutte le origini nell'ordine di input fino a quando non viene esaurita l'origine dati più piccola. Per utilizzare questa opzione è importante che i dati siano stati ordinati tramite un nodo Ordina.
- **Unione in base a un campo chiave**, come ad esempio *ID cliente*, per specificare il modo in cui associare i record da un'origine dati con i record da altre origini dati. Sono supportati diversi tipi di unioni, quali unione interna, unione esterna completa, unione esterna parziale e anti-unione. Per ulteriori informazioni, consultare l'argomento "Tipi di unioni".
- **Unione in base alla condizione** significa che l'unione verrà eseguita solo se viene soddisfatta la condizione specificata. È possibile specificare la condizione direttamente nel nodo, oppure crearla con il Generatore di espressioni.

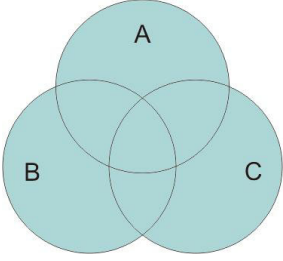
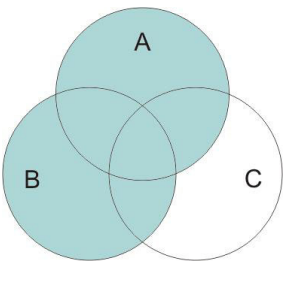
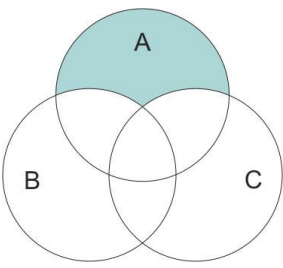
Tipi di unioni

Quando si utilizza un campo chiave per l'unione di dati, è utile considerare attentamente i record da escludere e quelli da includere. È disponibile una vasta gamma di unioni, illustrate di seguito in modo dettagliato.

I due tipi di unioni di base vengono definiti unioni interne e unioni esterne. Questi metodi vengono spesso utilizzati per unire le tabelle di insiemi di dati correlati in base ai valori comuni di un campo chiave, quale *ID cliente*. Le unioni interne consentono di ottenere un'unione corretta, con un insieme di dati di output comprendente solo record completi. Oltre a comprendere i record completi dei dati uniti, le unioni esterne consentono di includere dati univoci da una o più tabelle di input.

Di seguito viene fornita una descrizione più dettagliata dei tipi di unioni supportate.

	<p>Un'unione interna comprende solo i record in cui è presente un valore del campo chiave comune a tutte le tabelle di input. I record non corrispondenti non verranno quindi inclusi nell'insieme di dati di output.</p>
-------------------------------------------------------------------------------------	----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

	<p>Un'unione esterna completa comprende tutti i record, sia corrispondenti che non corrispondenti, delle tabelle di input. L'unione esterna sinistra e l'unione esterna destra, descritte di seguito, vengono definite unioni esterne parziali.</p>
	<p>Un'unione esterna parziale comprende tutti i record corrispondenti in base al campo chiave e i record non corrispondenti delle tabelle specificate. In altre parole, tutti i record di alcune tabelle e solo i record corrispondenti di altre. È possibile selezionare le tabelle da includere nell'unione esterna (per esempio, A e B della figura) utilizzando il pulsante Seleziona nella scheda Unione. Quando vengono unite soltanto due tabelle, le unioni parziali vengono denominate anche unione esterna sinistra o unione esterna destra. Poiché IBM SPSS Modeler supporta l'unione di più di due tabelle, in questo contesto viene utilizzata la definizione di unione esterna parziale.</p>
	<p>Un'anti-unione comprende soltanto i record non corrispondenti della prima tabella di input (tabella A della figura). Questo tipo di unione rappresenta l'opposto di un'unione interna e non determina l'inclusione di record completi nell'insieme di dati di output.</p>

Per esempio, se si hanno informazioni sulle aziende agricole in un insieme di dati e le richieste di risarcimento relative alle aziende agricole in un altro, è possibile abbinare i record della prima origine alla seconda origine utilizzando le opzioni di unione.

Per determinare se un cliente dell'azienda agricola campione ha presentato una richiesta di risarcimento, utilizzare l'opzione unione interna per restituire un elenco che mostra i punti di corrispondenza di tutti gli ID dei due campioni.

	id	name	region	farmsize	rainfall	landquality	farmincome	maincrop	claimtype	claimvalue
1	id604	name604	southwest	1860.000	103.0...	3.000	625251.000	potatoes	decomm...	281082.0...
2	id605	name605	north	1700.000	46.000	8.000	621148.000	wheat	decomm...	122006.0...
3	id620	name620	north	880.000	74.000	6.000	426988.000	rapeseed	arable_de	118885.0...

Figura 2. Esempio di output per un'integrazione di unione interna

L'utilizzo dell'opzione unione esterna completa restituisce tutti i record corrispondenti e non corrispondenti delle tabelle di input. Per i valori incompleti verrà utilizzato il valore mancante di sistema (\$null\$).

	id	name	region	farmsize	rainfall	landquality	farmincome	maincrop	claimtype	claimvalu
1	id601	\$null\$	\$null\$	\$null\$	\$null\$	\$null\$	\$null\$	\$null\$	decomm...	74703.1C
2	id602	name602	north	1780.000	42.000	9.000	734118.000	maize	\$null\$	\$nul
3	id604	name604	southwest	1860.000	103.0...	3.000	625251.000	potatoes	decomm...	281082.0
4	id605	name605	north	1700.000	46.000	8.000	621148.000	wheat	decomm...	122006.0
5	id606	\$null\$	\$null\$	\$null\$	\$null\$	\$null\$	\$null\$	\$null\$	arable_de	122135.0

Figura 3. Esempio di output per un'integrazione di unione esterna completa

Un'unione esterna parziale comprende tutti i record corrispondenti in base al campo chiave e i record non corrispondenti delle tabelle specificate. La tabella visualizza tutti i record corrispondenti del campo ID e i record corrispondenti del primo insieme di dati.

	id	claimtype	claimvalue	name	region	farmsize	rainfall	landquality	farmincome	maincrop
1	id602	\$null\$	\$null\$	name602	north	1780.000	42.000	9.000	734118.000	maize
2	id604	decomm...	281082.0...	name604	southwest	1860.000	103.0...	3.000	625251.000	potatoes
3	id605	decomm...	122006.0...	name605	north	1700.000	46.000	8.000	621148.000	wheat
4	id607	\$null\$	\$null\$	name607	southeast	1820.000	29.000	6.000	211605.000	maize
5	id608	\$null\$	\$null\$	name608	southeast	1640.000	108.0...	7.000	1167040.0...	maize
6	id609	\$null\$	\$null\$	name609	southwest	1600.000	101.0...	5.000	756755.000	wheat
7	id615	\$null\$	\$null\$	name615	midlands	920.000	86.000	6.000	442554.000	potatoes
8	id618	\$null\$	\$null\$	name618	southeast	1180.000	98.000	3.000	368646.000	maize

Figura 4. Esempio di output per un'integrazione di unione esterna parziale

Se viene utilizzata l'opzione anti-unione, la tabella restituisce soltanto i record non corrispondenti della prima tabella di input.

	id	name	region	farmsize	rainfall	landquality	farmincome	maincrop
1	id602	name602	north	1780.000	42.000	9.000	734118.000	maize
2	id607	name607	southeast	1820.000	29.000	6.000	211605.000	maize
3	id608	name608	southeast	1640.000	108.0...	7.000	1167040.0...	maize
4	id609	name609	southwest	1600.000	101.0...	5.000	756755.000	wheat
5	id615	name615	midlands	920.000	86.000	6.000	442554.000	potatoes
6	id618	name618	southeast	1180.000	98.000	3.000	368646.000	maize
7	id619	name619	north	840.000	64.000	8.000	457552.000	potatoes

Figura 5. Esempio di output per un'integrazione anti-unione

Definizione di chiavi e metodo di unione

La scheda Unione del nodo Unione contiene i campi riportati di seguito.

Metodo di unione. Selezionare **Ordine** o **Chiavi** per specificare il metodo di unione dei record. La selezione di **Chiavi** consente di attivare la parte inferiore della finestra di dialogo.

- **Ordine.** Unisce i record in base all'ordine in modo che il record *n* di ogni input venga unito per generare il record *n* di output. Quando sono stati generati tutti i record corrispondenti a un record di input, non vengono più creati record di output. Questo significa che il numero di record creati corrisponde al numero di record dell'insieme di dati più piccolo.
- **Chiavi.** Utilizza un campo chiave, per esempio *ID transazione*, per unire record con lo stesso valore del campo chiave. Questa opzione è equivalente a una "unione di uguaglianza" (equi-join) di database. Se un valore chiave ricorre più volte, verranno restituite tutte le possibili combinazioni. Per esempio, se record con lo stesso valore di campo chiave *A* contengono valori *B*, *C* e *D* diversi in altri campi, i campi uniti genereranno un record separato per ogni combinazione di *A* con il valore *B*, *A* con il valore *C* e *A* con il valore *D*.

Nota: nel metodo di unione in base alla chiave, i valori null non vengono considerati identici e non verranno inclusi nell'unione.

- **Condizione.** Utilizzare questa opzione per specificare una condizione per l'unione. Per ulteriori informazioni, consultare l'argomento "Specifica delle condizioni per un'unione" a pagina 77.

Chiavi possibili. Elenca solo i campi con nomi esattamente corrispondenti in tutte le origini dati di input. Selezionare un campo dall'elenco e utilizzare il pulsante con la freccia per aggiungerlo come campo chiave da utilizzare per l'unione dei record. È possibile utilizzare più campi chiave. I campi di input non corrispondenti si possono rinominare tramite un nodo Filtro o dalla scheda Filtro di un nodo origine.

Chiavi per l'unione. Elenca tutti i campi utilizzati per unire i record da tutte le origini dati di input in base ai valori dei campi chiave. Per rimuovere una chiave dall'elenco, selezionarla e utilizzare il pulsante con la freccia per includerla nuovamente nell'elenco Chiavi possibili. L'opzione seguente è attivata solo quando vengono selezionati più campi.

Unisci campi chiave duplicati. Quando è selezionato più di un campo chiave, questa opzione assicura che sia presente un solo campo di output con tale nome. Tale opzione è attivata per default, a meno che i flussi siano stati importati da versioni precedenti di IBM SPSS Modeler. Quando questa opzione è disattivata, i campi chiave duplicati devono essere rinominati o esclusi utilizzando la scheda Filtro nella finestra di dialogo del nodo Unione.

Includi solo record corrispondenti (unione interna). Selezionare per unire solo record completi.

Includi record corrispondenti e non corrispondenti (unione esterna completa). Selezionare questa opzione per eseguire una "unione esterna completa". In questo caso, se i valori del campo chiave non sono presenti in tutte le tabelle di input, i record incompleti vengono comunque mantenuti. Il valore non definito (\$null\$) viene aggiunto al campo chiave e incluso nel record di output.

Includi record non corrispondenti selezionati e corrispondenti (unione esterna parziale). Selezionare questa opzione per eseguire una "unione esterna parziale" delle tabelle selezionate in una finestra di dialogo secondaria. Fare clic su **Seleziona** per specificare le tabelle i cui record incompleti verranno mantenuti nell'unione.

Includi record nel primo insieme di dati senza corrispondenza (anti-unione). Selezionare questa opzione per eseguire un tipo di "anti-unione" in cui vengono passati a valle soltanto i record non corrispondenti del primo insieme di dati. È possibile specificare l'ordine dei dataset di input utilizzando le frecce disponibili nella scheda Input. Con questo tipo di unione, nell'insieme di dati di output non vengono inclusi record completi. Per ulteriori informazioni, consultare l'argomento "Tipi di unioni" a pagina 74.

Selezione di dati per unioni parziali

Per eseguire un'unione esterna parziale, è necessario selezionare una o più tabelle di cui verranno mantenuti i record incompleti. Può per esempio rivelarsi utile mantenere tutti i record di una tabella dei clienti, ma soltanto i record corrispondenti della tabella dei mutui ipotecari.

Colonna Unione esterna. Nella colonna *Unione esterna*, selezionare gli insiemi di dati da includere interamente. Nel caso di un'unione parziale, per gli insiemi di dati selezionati in questo contesto verranno mantenuti sia i record sovrapposti che i record incompleti. Per ulteriori informazioni, consultare l'argomento "Tipi di unioni" a pagina 74.

Specifica delle condizioni per un'unione

Impostando il metodo di unione su **Condizione**, è possibile specificare una o più condizioni che devono essere soddisfatte perché l'unione possa essere eseguita.

È possibile inserire le condizioni direttamente nel campo Condizione oppure crearle con l'aiuto del Generatore di espressioni facendo clic sul simbolo del calcolatore alla destra del campo.

Filtro di campi dal nodo Unione

I nodi Unione includono un comodo metodo per filtrare o rinominare i campi duplicati risultanti dall'unione di più sorgenti dati. Fare clic sulla scheda **Filtro** nella finestra di dialogo per selezionare le opzioni di filtro.

Le opzioni illustrate di seguito sono quasi identiche a quelle disponibili per il nodo Filtro. Nel menu Filtro sono tuttavia disponibili opzioni aggiuntive che non verranno illustrate di seguito. Per ulteriori informazioni, consultare l'argomento "Applicazione di filtri o ridenominazione di campi" a pagina 122.

Campo. Visualizza i campi di input dalle origini dati attualmente connesse.

Tag. Elenca il nome (o il numero) del tag associato al collegamento dell'origine dati. Fare clic sulla scheda **Input** per modificare i collegamenti attivi per questo nodo Unione.

Nodo origine. Visualizza il nodo di origine di cui vengono uniti i dati.

Nodo connesso. Visualizza il nome del nodo connesso al nodo Unione. Spesso il data mining complesso richiede diverse operazioni di unione o accodamento che possono includere lo stesso nodo origine. Il nome del nodo connesso consente di operare una distinzione.

Filtro. Visualizza le connessioni correnti tra il campo di input e di output. Le connessioni attive sono contrassegnate da una freccia continua. Le connessioni contrassegnate da una X rossa indicano i campi filtrati.

Campo. Elenca i campi di output dopo l'unione o l'accodamento. I campi duplicati vengono visualizzati in rosso. Fare clic sul campo Filtro per disattivare i campi duplicati.

Visualizza campi correnti. Selezionare per visualizzare le informazioni relative ai campi selezionati per essere utilizzati come campi chiave.

Visualizza impostazioni campi inutilizzati. Selezionare per visualizzare le informazioni relative ai campi non attualmente in uso.

Impostazione di tag e ordine dei dati di input

Mediante la scheda Input delle finestre di dialogo dei nodi Unione e Accodamento è possibile specificare l'ordine delle sorgenti dati di input e apportare qualsiasi modifica al nome di tag per ogni sorgente.

Tag e ordine dei dataset di input. Selezionare per unire oppure accodare solo record completi.

- **Tag.** Elenca i nomi di tag correnti per ciascuna origine dati di input. I nomi di tag o **tag** rappresentano un metodo per identificare in modo univoco i collegamenti dei dati per le operazioni di unione o accodamento. Per comprendere il funzionamento del flusso di dati in Clementine, si può immaginare una serie di tubi che trasportano acqua e che a un certo punto si uniscono in una sola condotta. Analogamente, in IBM SPSS Modeler i dati seguono percorsi separati e vengono uniti in un punto preciso, che costituisce spesso un'interazione complessa tra le diverse sorgenti dati. I tag rappresentano un mezzo per la gestione degli input (paragonabili ai tubi dell'acqua) in un nodo Unione o Accodamento, in modo che, se il nodo viene salvato o disconnesso, i collegamenti vengano mantenuti e siano facilmente identificabili.

Quando si connettono sorgenti dati aggiuntive a un nodo Unione o Accodamento, vengono creati automaticamente tag di default che utilizzano numeri per rappresentare l'ordine di connessione dei nodi. Tale ordine non è correlato all'ordine dei campi negli insiemi di dati di input o output. È possibile modificare il tag di default specificando un nuovo nome nella colonna *Tag*.

- **Nodo origine.** Visualizza il nodo di origine di cui vengono combinati i dati.
- **Nodo connesso.** Visualizza il nome del nodo connesso al nodo Unione o Accodamento. Spesso il data mining complesso richiede diverse operazioni di unione che possono includere lo stesso nodo origine. Il nome del nodo connesso consente di operare una distinzione.
- **Campi.** Elenca il numero di campi in ciascuna origine dati.

Visualizza tag correnti. Selezionare per visualizzare i tag utilizzati in modo attivo dal nodo Unione o Accodamento. In altri termini, i tag correnti identificano i collegamenti al nodo che includono il flusso di dati. Per tornare all'immagine dei tubi, i tag correnti sono analoghi ai tubi in cui l'acqua passa attualmente.

Visualizza impostazioni tag inutilizzati. Selezionare per visualizzare i tag o i collegamenti precedentemente utilizzati per la connessione al nodo Unione o Accodamento, ma attualmente non connessi ad un'origine dati. Questi tag possono essere paragonati a tubi vuoti e inutilizzati in un sistema di tubature. È possibile scegliere se connettere questi "tubi" a una nuova sorgente oppure rimuoverli. Per rimuovere i tag inutilizzati dal nodo, fare clic su **Cancella**. In questo modo, tutti i tag inutilizzati verranno rimossi con un'unica operazione.

Impostazioni di ottimizzazione dell'unione

Il sistema fornisce due opzioni che consentono di unire i dati con maggior efficacia in alcune situazioni. Queste opzioni consentono di ottimizzare l'unione quando un dataset di input è significativamente più grande degli altri insiemi di dati oppure quando i dati sono già ordinati in base a tutti o ad alcuni campi chiave utilizzati per l'unione.

Un dataset di input è relativamente grande. Selezionare questa opzione per indicare che uno dei dataset di input è molto più grande degli altri. Il sistema può memorizzare nella cache gli insiemi di dati più piccoli e quindi eseguire l'unione elaborando l'insieme di dati grande senza memorizzarlo nella cache o ordinarlo. Questo tipo di unione viene normalmente utilizzato con i dati presentati mediante uno schema a stella o un disegno analogo, dove vi è una grande tabella centrale di dati condivisi (è il caso, per esempio, dei dati transazionali). Se si seleziona questa opzione, fare clic su **Seleziona** per specificare l'insieme di dati grande. Si noti che è possibile selezionare solo *un* insieme di dati grande. La seguente tabella riassume le unioni che è possibile ottimizzare con questo metodo.

Tabella 11. Riepilogo delle ottimizzazioni di unione

Tipo di unione	Possibilità di ottimizzazione per un dataset di input grande
Inner	Sì
Parziale	Sì, se l'insieme di dati grande non presenta record incompleti.
Full	No
Anti-unione	Sì, se l'insieme di dati grande è il primo input.

Tutti gli input sono già ordinati in base ai campi chiave. Selezionare questa opzione per indicare che i dati di input sono già ordinati in base a uno o più dei campi chiave utilizzati per l'unione. Assicurarsi che *tutti* i dataset di input siano ordinati.

Specifica criterio di ordinamento esistente. Specificare i campi già ordinati. Utilizzando la finestra di dialogo Seleziona campi, aggiungere i campi all'elenco. È possibile effettuare la selezione solo dai campi chiave utilizzati per l'unione (specificati nella scheda Unione). Nella colonna *Ordine*, specificare se ciascun campo è ordinato in ordine crescente o decrescente. Se si specificano più campi, assicurarsi di elencarli secondo l'ordinamento corretto. Utilizzare le frecce a destra dell'elenco per disporre i campi nell'ordine corretto. In caso di errore nell'indicazione del criterio di ordinamento esistente corretto, al momento dell'esecuzione del flusso verrà visualizzato un errore che indica il numero dei record nei quali l'ordinamento è incoerente rispetto all'impostazione specificata.

A seconda della distinzione tra caratteri maiuscoli/minuscoli del metodo di classificazione utilizzato dal database, è possibile che l'ottimizzazione non funzioni correttamente nel caso in cui uno o più input vengano ordinati dal database. Per esempio, se si hanno due input e soltanto in uno dei due viene fatta distinzione tra caratteri maiuscoli/minuscoli, i risultati dell'ordinamento potrebbero essere diversi. A causa dell'ottimizzazione dell'unione, i record vengono elaborati in base al relativo ordinamento. Per questa ragione, se gli input vengano ordinati utilizzando metodi di classificazione diversi, il nodo Unione

segnala un errore e visualizza il numero di record in cui l'ordinamento è incoerente. Se tutti gli input provengono da un'unica sorgente o sono ordinati utilizzando classificazioni reciprocamente inclusive, i record possono essere uniti correttamente.

Nota: la velocità di unione può beneficiare dell'attivazione dell'elaborazione parallela.

Nodo Accodamento

È possibile utilizzare i nodi Accodamento per concatenare insieme di record. A differenza dei nodi Unione, che consentono di unire record da sorgenti diverse, i nodi Accodamento leggono e passano a valle tutti i record di un'unica sorgente. Vengono quindi letti i record della sorgente successiva utilizzando la stessa struttura di dati della prima o principale sorgente di input (numero di record, di campi e così via). Quando la sorgente principale include più campi rispetto a un'altra origine di input, per tutti i valori incompleti verrà utilizzato il valore stringa null di sistema (\$null\$).

I nodi Accodamento possono essere utilizzati per combinare insieme di dati con strutture simili ma dati diversi. È possibile infatti che i dati delle transazioni siano archiviati in file diversi corrispondenti a periodi di tempo diversi, per esempio in un file di dati relativi alle vendite di marzo e in un file diverso relativo alle vendite di aprile. Presupponendo che questi file utilizzino la stessa struttura (ovvero gli stessi campi nello stesso ordine), il nodo Accodamento consente di eseguire l'unione dei file in un unico file di dimensioni maggiori che sarà quindi possibile analizzare.

Nota: per accodare i file, i livelli di misurazione dei campi devono essere simili. Per esempio, a un campo *Nominale* non può essere accodato un campo con livello di misurazione *Continuo*.

Impostazione delle opzioni relative al nodo Accodamento

Corrispondenze campi in base a. Selezionare un metodo da utilizzare per la corrispondenza tra i campi che si desidera accodare.

- **Posizione.** Selezionare questa opzione per accodare insieme di dati in base alla posizione dei campi nell'origine dati principale. Quando si utilizza questo metodo, è necessario che i dati siano ordinati in modo da garantire un accodamento appropriato.
- **Nome.** Selezionare questa opzione per accodare insieme di dati in base al nome dei campi nei dataset di input. Selezionare inoltre **Caratteri maiuscoli/minuscoli** per attivare la distinzione tra caratteri maiuscoli/minuscoli quando si esegue la corrispondenza tra i nomi dei campi.

Campo di output. Elenca i nodi origine connessi al nodo Accodamento. Il primo nodo nell'elenco è l'origine di input principale. È possibile ordinare i campi nell'elenco visualizzato facendo clic sull'intestazione della colonna. Questa impostazione, tuttavia, non consente di riordinare i campi nel dataset.

Includi campi da. Selezionare **Solo dataset principale** per creare i campi di output in base ai campi del dataset principale. Il dataset principale rappresenta il primo input, specificato nella scheda Input. Selezionare **Tutti gli insiemi di dati** per creare campi di output per tutti i campi di tutti gli insiemi di dati, a prescindere dalla presenza di un campo corrispondente in tutti i dataset di input.

Contrassegna i record con insieme dei dati di origine nel campo. Selezionare questa opzione per aggiungere un campo supplementare al file di output i cui valori indicano il dataset di origine per ogni record. Specificare un nome nel campo testo. Il nome di campo predefinito è *Input*.

Nodo Elimina duplicati

I record duplicati di un insieme di dati devono essere rimossi prima di iniziare il data mining. Ad esempio, in un database di marketing, è possibile che gli utenti vengano visualizzati più volte con indirizzo o dati della società diversi. È possibile utilizzare il nodo Distinto per individuare, o rimuovere, i record duplicati nel dataset o creare un singolo record composito dai vari record.

Utilizzando il nodo Elimina duplicati, è possibile rimuovere record duplicati passando il primo record distinto nel flusso di dati oppure è possibile trovare i record duplicati scartando il primo record e passando nel flusso tutti i duplicati.

Inoltre, è possibile definire un criterio di ordinamento all'interno dei singoli valori di chiavi distinte per i risultati restituiti. Se si desidera che venga restituita una riga specifica per ogni valore di chiave distinta, è necessario ordinare i record all'interno del nodo Elimina duplicati anziché utilizzare un nodo Ordina a monte (vedere "Ordinamento dei record nel nodo Elimina duplicati" di seguito).

Modalità. Specificare se creare un record composito o se includere o escludere (scartare) il primo record.

- **Crea un record composito per ogni gruppo.** Fornisce un modo per aggregare i campi non numerici; selezionando questa opzione si rende disponibile la scheda Composito su cui è possibile specificare come creare i record composti. Consultare "Impostazioni di Composito distinto" a pagina 83 per ulteriori informazioni.
- **Includi solo primo record in ogni gruppo.** Include il primo record distinto nel flusso di dati e rimuove eventuali duplicati.
- **Scarta solo il primo record in ogni gruppo.** Scarta il primo record distinto trovato e passa tutti i record duplicati nel flusso di dati. Questa opzione può essere utilizzata per *trovare* duplicati nei dati in modo da analizzarli in seguito all'interno del flusso.

Campi chiave da raggruppare. Elenca il campo o i campi utilizzati per stabilire se i record sono identici. È possibile:

- Aggiungere campi all'elenco utilizzando il pulsante di selezione dei campi a destra.
- Eliminare i campi dall'elenco utilizzando il pulsante X (rimuovi) rosso.

All'interno dei gruppi, ordina record per. Elenca i campi utilizzati per stabilire in che modo i record vengono ordinati all'interno di ogni valore di chiave distinta e se questi sono in ordine crescente o decrescente. È possibile:

- Aggiungere campi all'elenco utilizzando il pulsante di selezione dei campi a destra.
- Eliminare i campi dall'elenco utilizzando il pulsante X (rimuovi) rosso.
- Spostare i campi utilizzando i pulsanti su o giù, se si esegue l'ordinamento in base a più campi.

Criterio di ordinamento di default. Specificare se, per default, i record vengono disposti in ordine **Crescente** o **Decrescente**.

Ordinamento dei record nel nodo Elimina duplicati

Utilizzando l'opzione **All'interno dei gruppi, ordina record per** all'interno del nodo Elimina duplicati è possibile restituire una riga specifica per ogni valore di chiave distinta, senza che sia necessario utilizzare un nodo Ordina precedente. Per esempio, si supponga di disporre dei seguenti dati relativi alle età dei consumatori di farmaci prescritti.

Tabella 12. Dati utente cura prescritta.

Età	Cura
50	Cura A
71	Cura B
44	Cura A
65	Cura X
39	Cura A
75	Cura C
72	Cura Y

Tabella 12. Dati utente cura prescritta (Continua).

Età	Cura
57	Cura X
79	Cura Y
69	Cura C
74	Cura B
85	Cura Y
69	Cura X

Per trovare il consumatore più anziano per ogni cura, è necessario specificare l'inserimento del primo record di ogni gruppo, utilizzare Cura come campo chiave e utilizzare Età come campo di ordinamento, impostato su ordine Decrescente. L'ordine di input non ha alcun effetto sul risultato in quanto la selezione dell'ordinamento specifica quale delle molte righe di una data Cura deve essere restituita. L'output di dati finale sarebbe il seguente.

Tabella 13. Dati utente cura prescritta ordinati.

Età	Cura
50	Cura A
74	Cura B
75	Cura C
69	Cura X
85	Cura Y

Impostazioni di ottimizzazione del nodo Elimina duplicati

Se i dati su cui si lavora presentano solo un numero ridotto di record o sono già stati ordinati, è possibile ottimizzarne la gestione per consentire a IBM SPSS Modeler di elaborarli in modo più efficiente.

Nota: se viene selezionata l'opzione **Il dataset di input presenta un numero ridotto di chiavi distinte** oppure viene utilizzata la generazione SQL per il nodo, possono essere restituite tutte le righe all'interno del valore della chiave distinta; per controllare quale riga viene restituita all'interno di una chiave distinta, è necessario specificare il criterio di ordinamento utilizzando i campi **All'interno dei gruppi, ordina record per** nella scheda Impostazioni. Le opzioni di ottimizzazione non influiscono sull'output dei risultati del nodo Elimina duplicati, a condizione che nella scheda Impostazioni sia stato specificato un criterio di ordinamento.

Il dataset di input presenta un numero ridotto di chiavi distinte. Selezionare questa opzione se i record e/o i valori univoci dei campi chiave sono in numero ridotto. In questo modo si migliorano le prestazioni.

Il dataset di input è già ordinato per campi di raggruppamento e campi di ordinamento definiti nella scheda Impostazioni. Selezionare questa opzione solo se i dati sono già ordinati in base a tutti i campi elencati nella scheda Impostazioni in **All'interno dei gruppi, ordina record per** e se il criterio di ordinamento dei dati è lo stesso, crescente o decrescente. In questo modo si migliorano le prestazioni.

Disattiva generazione SQL. Selezionare questa opzione per disattivare la generazione SQL relativa al nodo.

Impostazioni di Composito distinto

Se i dati su cui si sta lavorando presentano più record, ad esempio per la stessa persona è possibile ottimizzare il modo in cui vengono gestiti i dati creando un unico record composito, o aggregato, da elaborare. Se è installato IBM SPSS Modeler Entity Analytics è possibile utilizzarlo per combinare, o appiattare, record duplicati che sono l'output di SPSS Entity Analytics.

Nota: Questa scheda è disponibile solo quando si seleziona **Crea un record composito per ogni gruppo** nella scheda Impostazioni.

Ad esempio, SPSS Entity Analytics contrassegna tre record come medesima entità, come illustrato nella seguente tabella.

Tabella 14. Esempio di più record per la stessa entità.

\$EA-ID	Nome	Età	Banca	Istruzione superiore	Debito totale
0003	Bob Jones	27	K	Scuola	27000
0003	Robert Jones	35	A	Grado	42000
0003	Robbie Jones	27	D	Laurea	7000

Si desidera aggregare questi tre record in un singolo record da utilizzare downstream. Si potrebbe utilizzare il nodo Aggregazione per aggiungere il debito totale e calcolare un'età media; tuttavia, non è possibile effettuare una media dei dettagli come i nomi, le banche e così via. Se si specifica quali dettagli utilizzare per creare un record composito è possibile ricavare un singolo record.

Dalla tabella si potrebbe creare un record composito scegliendo i seguenti dettagli.

- Per **Nome** utilizzare il primo record
- Per **Età** utilizzare la più alta
- Per **Banca** concatenare tutti i valori senza delimitatore
- Per **Istruzione superiore** utilizzare il primo dell'elenco (laurea)
- Per **Debito** utilizzare il totale

Combinando (o aggregando) questi dettagli si ottiene un singolo record composito che contiene i seguenti dettagli.

- Nome: Bob Jones
- Età: 35
- Banca KND
- Istruzione superiore: Laurea
- Debito: 76000

Questo fornisce un'immagine di Bob Jones, una persona laureata, che ha almeno 35 anni, con almeno tre conti in banca noti e un debito totale consistente.

Impostazione delle opzioni relative alla scheda Composito

Campo. Questa colonna visualizza tutti i campi, tranne i campi chiave del modello dati, nel loro ordine naturale; se il nodo non è connesso non vengono visualizzati campi. Per ordinare le righe alfabeticamente in base al nome del campo, fare clic sull'intestazione della colonna. È possibile selezionare più di una riga utilizzando Maiusc-clic o Ctrl-clic. Inoltre, se si fa clic con il tasto destro del mouse sul campo viene visualizzato un menu da cui è possibile scegliere di selezionare tutte le righe, ordinare le righe in base al nome o al valore del campo con ordinamento ascendente o discendente, selezionare i campi in base alla

misurazione o al tipo di archiviazione oppure selezionare un valore per aggiungere automaticamente la stessa voce **Riempire con valori basati su** ad ogni riga selezionata.

Riempire con valori basati su. Selezionare il tipo di valore da utilizzare per il record composito per il **Campo**. Le opzioni disponibili dipendono dal tipo di campo.

- Per i campi con intervalli numerici è possibile scegliere tra:
 - Primo record nel gruppo
 - Ultimo record nel gruppo
 - Totale
 - Media
 - Minimo
 - Massimo
 - Personalizzato
- Per i campi i data o ora è possibile scegliere tra:
 - Primo record nel gruppo
 - Ultimo record nel gruppo
 - Primo
 - Più recente
 - Personalizzato
- Per campi stringa o senza tipo è possibile scegliere tra:
 - Primo record nel gruppo
 - Ultimo record nel gruppo
 - Primo alfanumerico
 - Ultimo alfanumerico
 - Personalizzato

In ciascun caso è possibile utilizzare l'opzione **Personalizzato** per esercitare un maggior controllo su quale valore viene utilizzato per costituire il record composito. Consultare “Scheda Composito distinto - Personalizzato” per ulteriori informazioni.

Includi conteggio record nel campo. Selezionare questa opzione per includere un campo aggiuntivo in ogni record di output, denominato Numero_record per default. Tale campo indica il numero dei record di input che sono stati aggregati per formare ogni record aggregato. Per creare un nome personalizzato per questo campo immettere la voce nel campo di modifica.

Scheda Composito distinto - Personalizzato

La finestra di dialogo Riempimento personalizzato fornisce un maggiore controllo su quale valore viene utilizzato per completare il nuovo record composito. Notare che è necessario prima istanziare i dati per poter utilizzare questa opzione se si sta personalizzando una singola riga Campo nella scheda Composito.

Nota: Questa finestra di dialogo è disponibile solo quando si seleziona il valore Personalizzato nella colonna **Riempire con valori basati su** nella scheda Composito.

A seconda del tipo di campo, è possibile scegliere una delle seguenti opzioni.

- **Seleziona per frequenza.** Scegliere un valore in base alla frequenza con cui ricorre nel record di dati.

Nota: Non disponibile per i campi con tipo Continuo, Senza tipo o Data/Ora.

- **Utilizza.** Selezionare Più frequente o Meno frequente.
- **Pari merito.** Se esistono due o più record con la stessa frequenza di ricorrenza, specificare come selezionare il record richiesto. È possibile scegliere tra quattro opzioni: Utilizza primo, Utilizza ultimo, Utilizza più basso o Utilizza più alto.

- **Include il valore (T/F).** Selezionare questa opzione per convertire un campo in un indicatore che identifica se un record di un gruppo ha un valore specificato. È possibile quindi selezionare il **Valore** dall'elenco di valori relativi al campo selezionato.

Nota: Non disponibile se si seleziona più di una riga Campo nella scheda Composito

- **Prima corrispondenza nell'elenco.** Selezionare questa opzione dare priorità ad un valore da assegnare al record composito. È possibile quindi selezionare uno degli **Elementi** dall'elenco di quelli relativi al campo selezionato.

Nota: Non disponibile se si seleziona più di una riga Campo nella scheda Composito

- **Concatena i valori.** Selezionare questa opzione per conservare tutti i valori di un gruppo concatenandoli in una stringa. È necessario specificare un delimitatore da utilizzare tra ogni valore.

Nota: Questa è l'unica opzione disponibile se si selezionano una o più righe Campo con tipo Continuo, Senza tipo o Data/Ora.

- **Utilizza delimitatore.** È possibile scegliere di utilizzare uno **Spazio** o una **Virgola** come valore delimitatore nella stringa concatenata. In alternativa, nel campo **Altro** è possibile immettere un proprio carattere come valore delimitatore.

Nota: Disponibile solo se si seleziona l'opzione **Concatena i valori**.

Nodo Streaming TS

Il nodo Streaming TS (Time Series) può essere utilizzato per creare modelli di serie temporali e calcolarne i punteggi in un unico passo, e può essere utilizzato con o senza un nodo upstream Intervalli di tempo. Un modello di serie temporali separato viene creato per ogni campo obiettivo, tuttavia i nugget del modello non vengono aggiunti alla tavolozza dei modelli generati e non è possibile esaminare le informazioni sul modello.

Non è sempre necessario utilizzare un nodo Intervalli di tempo con il nodo Streaming TS, a patto che siano soddisfatti i seguenti punti:

- I dati devono essere distanziati uniformemente.
- I dati dispongono di etichette di serie temporali.
- Non è necessario aggiungere ulteriori record.
- I dati hanno una *periodicità* standard, ad esempio cinque giorni a settimana o otto ore al giorno.

I metodi di modellazione dei dati di serie temporali richiedono un intervallo uniforme tra una misurazione e l'altra, con gli eventuali valori mancanti indicati da righe vuote. Se i dati non soddisfano già questo requisito, un nodo upstream Intervalli di tempo può trasformare i valori in base alle esigenze. Per ulteriori informazioni, consultare l'argomento "Nodo Intervalli di tempo" a pagina 151.

Altri aspetti da evidenziare relativamente ai nodi Serie temporali sono riportati di seguito:

- I campi devono essere numerici.
- I campi data non possono essere utilizzati come input.
- Le partizioni vengono ignorate.

Il nodo Streaming TS stima i modelli di livellamento esponenziale, i modelli ARIMA (Autoregressive Integrated Moving Average, autoregressivi integrati a media mobile) univariati e i modelli ARIMA (o a funzione di trasferimento) multivariati per le serie temporali e genera previsioni basate sui dati delle serie temporali. È inoltre disponibile un Expert Modeler, che tenta di identificare e stimare automaticamente il modello ARIMA o il modello di livellamento esponenziale più adatto per uno o più campi obiettivo. Per ulteriori informazioni sui modelli di serie temporali, consultare l'argomento *Modelli di serie temporali* nel documento *IBM SPSS Modeler 16 - Nodi di modellazione*.

Il nodo Streaming TS è supportato per essere utilizzato in un ambiente di distribuzione in streaming, tramite IBM SPSS Modeler Solution Publisher, utilizzando il servizio di calcolo del punteggio di IBM SPSS Collaboration and Deployment Services o IBM InfoSphere Warehouse.

Opzioni dei campi di Streaming TS

Nella scheda Campi si indicano i campi da utilizzare nella creazione del modello. Per poter generare un modello, è necessario prima specificare i campi da utilizzare come obiettivi e come input. In genere il nodo Streaming TS utilizza le informazioni dei campi di un nodo Tipo upstream. Se si utilizza un nodo Tipo per selezionare i campi obiettivo e di input, non è necessario cambiare nessuna delle impostazioni presenti in questa scheda.

Utilizza impostazioni nodo Tipo. Questa opzione indica al nodo di utilizzare le informazioni sui campi da un nodo Tipo a monte. Questa è l'opzione di default.

Utilizzo delle impostazioni personalizzate. Questa opzione indica al nodo di utilizzare le informazioni sui campi specificate qui al posto di quelle date in un qualsiasi nodo Tipo a monte. Dopo avere selezionato questa opzione, specificare i campi nell'area sottostante. Si noti che i campi archiviati come date non vengono accettati né come campi obiettivo, né come campi di input.

- **Obiettivi.** Selezionare uno o più campi obiettivo. Questa operazione è simile all'impostazione del ruolo di un campo su *Obiettivo* in un nodo Tipo. I campi obiettivo dei modelli di serie temporali devono avere un livello di misurazione *Continuo*. Per ogni campo obiettivo viene creato un modello separato. Un campo obiettivo considera tutti i campi *Input* specificati come possibili input, tranne se stesso. Pertanto, lo stesso campo può essere incluso in entrambi gli elenchi; tale campo sarà utilizzato come possibile input per tutti i modelli tranne per quello in cui rappresenta il campo obiettivo.
- **Input.** Selezionare i campi input. Questa operazione è simile all'impostazione del ruolo di un campo su *Input* in un nodo Tipo. I campi di input per i modelli di serie temporali devono essere numerici.

Opzioni del modello Streaming TS

Metodo. Le opzioni disponibili sono Expert Modeler, Livellamento esponenziale o ARIMA. Selezionare **Criteri** per specificare le opzioni per il metodo selezionato.

- **Expert Modeler.** Scegliere questa opzione per utilizzare Expert Modeler, che individua automaticamente il modello con il miglior adattamento per ciascuna serie.
- **Livellamento esponenziale.** Utilizzare questa opzione per specificare un modello di livellamento esponenziale personalizzato.
- **ARIMA.** Utilizzare questa opzione per specificare un modello ARIMA personalizzato.

Crea campi di confidenza superiore e inferiore dei limiti. Selezionare questa opzione per generare campi per intervalli di confidenza per i valori stimati del modello e le autocorrelazioni dei residui. Quando è selezionato, il controllo **Larghezza limite di confidenza (%)** è abilitato.

Larghezza limite di confidenza (%). Specificare gli intervalli di confidenza da calcolare. È possibile specificare qualsiasi valore positivo inferiore a 100. Per impostazione predefinita, si utilizza un intervallo di confidenza al 95%.

Opzioni degli intervalli di tempo

In questa sezione della finestra di dialogo, è possibile scegliere se utilizzare le specifiche per le stime e le previsioni di un nodo upstream Intervalli di tempo o specificare queste impostazioni sul nodo Streaming TS.

Utilizza impostazioni di nodo Intervalli di tempo. Selezionare questa opzione per utilizzare le informazioni relative alle specifiche per le stime e le previsioni effettuate su un nodo upstream Intervalli di tempo.

La prima riga indica gli eventuali record esclusi dal modello o utilizzati come record di holdout.

La seconda riga fornisce informazioni sugli eventuali periodi di previsione specificati sul nodo Intervalli di tempo.

La dicitura **Nessun intervallo di tempo definito** visualizzata sulla prima riga indica che non è connesso alcun nodo Intervalli di tempo. Questa situazione genererà un errore nel tentativo di eseguire il flusso. Se si seleziona **Utilizza impostazioni di nodo Intervalli di tempo** è necessario includere un nodo Intervalli di tempo upstream dal nodo Streaming TS.

Specifica impostazioni. Selezionare questa opzione per specificare le impostazioni per le stime e le previsioni se non è presente un nodo Intervalli di tempo upstream dal nodo Streaming TS.

Stima fino a. Selezionare **Ultima osservazione valida** per stimare il modello utilizzando tutti i record fino all'ultima osservazione valida. I record utilizzati per la previsione iniziano dalla prima osservazione non valida. Le osservazioni *Valide* sono quelle in cui tutti i campi obiettivo contengono valori non null. Selezionare **Conteggio indietro dall'ultimo record** per riservare una serie dei record più recenti per la previsione. I record più recenti potrebbero avere valori per i campi obiettivo, sebbene questo non sia necessario. Se **Conteggio indietro dall'ultimo record** è selezionato, il controllo **Offset** è abilitato, dove è possibile specificare il numero di record da utilizzare solo per la previsione. Il numero di record di default è quattro, e il numero massimo è 99.

Nota: I record da utilizzare per la previsione devono contenere i valori in tutti i campi di input.

Intervallo di tempo. Specificare l'intervallo di tempo utilizzato per la stima e la previsione. Sono disponibili le seguenti opzioni: Anni, Trimestri, Mesi, Settimane, Giorni, Ore, Minuti e Secondi.

Criteri di Expert Modeler per le serie temporali

Tipo di modello. Sono disponibili le seguenti opzioni:

- **Tutti i modelli.** Expert Modeler considera entrambi i modelli ARIMA e di livellamento esponenziale.
- **Solo modelli di livellamento esponenziale.** Expert Modeler considera solo i modelli di livellamento esponenziale.
- **Solo modelli ARIMA.** Expert Modeler considera solo i modelli ARIMA.

Expert Modeler considera i modelli stagionali. Questa opzione viene attivata solo se è stata definita una periodicità per il file di dati attivo. Quando l'opzione è selezionata, Expert Modeler considera sia i modelli stagionali che non stagionali. Se l'opzione non è selezionata, Expert Modeler considera solo i modelli non stagionali.

Eventi e interventi. Consente di definire alcuni campi di input come campi evento o intervento. In questo modo, un campo viene identificato come contenente dati di serie temporali influenzati dagli eventi (situazioni ricorrenti prevedibili, quali promozioni sulle vendite) o dagli interventi (eventi occasionali quali black-out o sciopero dei dipendenti). Expert Modeler non considera le funzioni di trasferimento arbitrarie per gli input identificati come campi evento o intervento.

Per poter essere inclusi nell'elenco, i campi di input devono avere un livello di misurazione *Flag*, *Nominale* o *Ordinale* e devono essere numerici (per esempio, per un campo flag, 1/0 non True/False).

Valori anomali

Rileva automaticamente valori anomali. Per impostazione predefinita, il rilevamento automatico dei valori anomali non viene eseguito. Selezionare questa opzione per eseguire il rilevamento automatico dei valori anomali, quindi selezionare i tipi di valori anomali desiderati. Per ulteriori informazioni, consultare l'argomento "Gestione dei valori anomali" a pagina 91.

Criteri di livellamento esponenziale per le serie temporali

Tipo di modello. I modelli di livellamento esponenziale vengono classificati come stagionali o non stagionali¹. I modelli stagionali sono disponibili solo se la periodicità è stagionale. Le periodicità stagionali sono: periodi ciclici, anni, trimestri, mesi, giorni alla settimana, ore al giorno, minuti al giorno e secondi al giorno.

- **Semplice.** Questo modello è adatto per le serie che non evidenziano tendenza o stagionalità. L'unico parametro di livellamento rilevante è il livello. Il livellamento esponenziale semplice è molto simile a un modello ARIMA con zero ordini di autoregressione, un ordine di differenziazione, un ordine di media mobile e nessuna costante.
- **Tendenza lineare di Holt.** Questo modello è adatto per le serie che evidenziano una tendenza lineare e nessuna stagionalità. I parametri di livellamento rilevanti sono il livello e la tendenza e, in questo modello, non sono reciprocamente limitati dai rispettivi valori. Il modello di Holt è più generale di quello di Brown ma può richiedere più tempo per il calcolo delle stime nel caso di serie particolarmente estese. Il livellamento esponenziale di Holt è molto simile a un modello ARIMA con zero ordini di autoregressione, due ordini di differenziazione e due ordini di media mobile.
- **Tendenza lineare di Brown.** Questo modello è adatto per le serie che evidenziano una tendenza lineare e nessuna stagionalità. I parametri di livellamento rilevanti sono il livello e la tendenza, ma in questo modello tali parametri si presuppongono uguali. Il modello di Brown rappresenta quindi un caso speciale del modello di Holt. Il livellamento esponenziale di Brown è molto simile a un modello ARIMA con zero ordini di autoregressione, due ordini di differenziazione e due ordini di media mobile, con il coefficiente del secondo ordine di media mobile uguale alla metà del coefficiente del primo ordine al quadrato.
- **Tendenza smorzata.** Questo modello è adatto per le serie con una tendenza lineare in via di attenuazione e privo di stagionalità. I parametri di livellamento rilevanti sono il livello, la tendenza e la tendenza smorzata. Il livellamento esponenziale smorzato è molto simile a un modello ARIMA con un ordine di autoregressione, un ordine di differenziazione e due ordini di media mobile.
- **Stagionale semplice.** Questo modello è adatto per le serie prive di tendenza e con un effetto stagionale costante nel tempo. I parametri di livellamento rilevanti sono il livello e la stagione. Il livellamento esponenziale stagionale è molto simile a un modello ARIMA con zero ordini di autoregressione, un ordine di differenziazione, un ordine di differenziazione stagionale e gli ordini $1, p$ e $p+1$ di media mobile, in cui p è il numero dei periodi di un intervallo stagionale. Per i dati mensili, $p = 12$.
- **Additiva di Winters.** Questo modello è adatto per le serie con una tendenza lineare e un effetto stagionale costante nel tempo. I parametri di livellamento rilevanti sono il livello, la tendenza e la stagione. Il livellamento esponenziale additivo di Winters è molto simile a un modello ARIMA con zero ordini di autoregressione, un ordine di differenziazione, un ordine di differenziazione stagionale e gli ordini $p+1$ di media mobile, in cui p è il numero dei periodi di un intervallo stagionale. Per i dati mensili, $p = 12$.
- **Moltiplicativa di Winters.** Questo modello è adatto per le serie con una tendenza lineare e un effetto stagionale che cambia a seconda della grandezza della serie. I parametri di livellamento rilevanti sono il livello, la tendenza e la stagione. Il livello esponenziale moltiplicativo di Winters non è simile a nessun modello ARIMA.

Trasformazione obiettivo. È possibile specificare una trasformazione da eseguire su ogni variabile dipendente prima della modellazione. Per ulteriori informazioni consultare l'argomento *Trasformazioni di serie* nel documento *IBM SPSS Modeler 16 - Nodi di modellazione*.

- **Nessuna.** Non viene eseguita nessuna trasformazione.
- **Radice quadrata.** Viene eseguita la trasformazione radice quadrata.
- **Log naturale.** Viene eseguita la trasformazione logaritmica naturale.

1. Gardner, E. S. 1985. Exponential smoothing: The state of the art. *Journal of Forecasting*, 4, 1-28.

Criteri ARIMA per le serie temporali

Il nodo Serie temporale consente di creare modelli ARIMA stagionali o non stagionali personalizzati, noti anche come modelli di Box-Jenkins, con o senza un insieme di variabili di input (predittore) fisso ². È possibile definire funzioni di trasferimento per qualsiasi o tutte le variabili di input e specificare il rilevamento automatico dei valori anomali o di un insieme specifico di valori anomali.

Tutte le variabili di input specificate vengono incluse esplicitamente nel modello, al contrario di quanto avviene con l'utilizzo di Expert Modeler, in cui le variabili di input vengono incluse solo se hanno una relazione statisticamente significativa con la variabile obiettivo.

Modello

La scheda Modello consente di specificare la struttura di un modello ARIMA personalizzato.

Ordini ARIMA. Immettere i valori per i vari componenti ARIMA del modello nelle celle corrispondenti della griglia struttura. Tutti i valori devono essere interi non negativi. Per i componenti autoregressivo e media mobile, il valore rappresenta l'ordine massimo. Nel modello vengono inclusi tutti gli ordini inferiori positivi. Per esempio, se si specifica 2, il modello include gli ordini 2 e 1. Le celle della colonna Stagionale sono attivate solo se è stata definita una periodicità per l'insieme di dati attivo.

- **Autoregressivo (p).** Il numero di ordini autoregressivi nel modello. Gli ordini autoregressivi specificano quali valori precedenti della serie vengono utilizzati per prevedere i valori correnti. Per esempio, un ordine autoregressivo 2 specifica di utilizzare il valore dei due periodi precedenti della serie per prevedere il valore corrente.
- **Differenza (d).** Specifica l'ordine di differenziazione applicato alla serie prima di eseguire la stima dei modelli. La differenziazione è necessaria quando sono presenti delle tendenze (di norma, le serie che presentano delle tendenze sono non stazionarie e nei modelli ARIMA si presume che vi sia stazionarietà) e viene utilizzata per rimuoverne l'effetto. L'ordine di differenziazione corrisponde al grado di tendenza della serie, la differenziazione di primo grado tiene conto delle tendenze lineari, la differenziazione di secondo grado tiene conto delle tendenze quadratiche e così via.
- **Media mobile (q).** Il numero di ordini di media mobile nel modello. Gli ordini di media mobile specificano il modo in cui vengono utilizzate le deviazioni provenienti dalla media della serie per prevedere i valori correnti. Per esempio, gli ordini di media mobile 1 e 2 specificano di considerare le deviazioni dalla media della serie degli ultimi due periodi precedenti per prevedere i valori correnti della serie.

Gradi stagionali. I componenti autoregressivo, media mobile e differenziazione stagionali hanno lo stesso ruolo delle corrispettive controparti non stagionali. Per gli ordini stagionali tuttavia, i valori di serie correnti vengono influenzati dai valori di serie precedenti separati da uno o più periodi stagionali. Ad esempio, per i dati mensili (periodo stagionale di 12), un ordine stagionale 1 è il valore della serie corrente è influenzato dal valore della serie che precede di 12 periodi quello corrente. Specificare un ordine stagionale 1, per i dati mensili, è quindi come specificare un ordine non stagionale 12.

Trasformazione obiettivo. È possibile specificare una trasformazione da eseguire su ogni variabile obiettivo prima della modellazione. Per ulteriori informazioni consultare l'argomento *Trasformazioni di serie* nel documento *IBM SPSS Modeler 16 - Nodi di modellazione*.

- **Nessuna.** Non viene eseguita nessuna trasformazione.
- **Radice quadrata.** Viene eseguita la trasformazione radice quadrata.
- **Log naturale.** Viene eseguita la trasformazione logaritmica naturale.

2. Box, G. E. P., G. M. Jenkins, and G. C. Reinsel. 1994. *Time series analysis: Forecasting and control*, 3rd ed. Englewood Cliffs, N.J.: Prentice Hall.

Includi costante nel modello. Il processo di inclusione di una costante è standard a meno che non si abbia la certezza che il valore generale medio della serie sia 0. Quando si applica la differenziazione, si consiglia di escludere la costante.

Funzioni di trasferimento

La scheda Funzioni di trasferimento consente di definire le funzioni di trasferimento per uno o tutti i campi di input. Tali funzioni consentono di specificare le modalità con cui i valori passati dei campi vengono utilizzati per prevedere i valori futuri della serie obiettivo.

La scheda viene visualizzata solo se sono specificati campi di input (con il ruolo impostato su *Input*), nel nodo Tipo oppure nella scheda Campi del nodo Streaming TS (selezionare **Utilizza impostazioni personalizzate—Input**).

L'elenco in alto mostra tutti i campi di input. Le altre informazioni di questa finestra di dialogo sono specifiche del campo di input selezionato nell'elenco.

Ordini della funzione di trasferimento. Immettere i valori per i vari componenti della funzione di trasferimento nelle celle corrispondenti della griglia struttura. Tutti i valori devono essere interi non negativi. Per i componenti numeratore e denominatore, il valore rappresenta l'ordine massimo. Nel modello vengono inclusi tutti gli ordini inferiori positivi. Inoltre, per i componenti numeratore l'ordine 0 viene incluso sempre. Per esempio, se si indica 2 come numeratore, il modello include gli ordini 2, 1 e 0. Se si indica 3 come denominatore, il modello include gli ordini 3, 2 e 1. Le celle della colonna Stagionale sono attivate solo se è stata definita una periodicità per l'insieme di dati attivo.

Numeratore. L'ordine del numeratore della funzione di trasferimento indica quali valori precedenti della serie indipendente (predittore) selezionata vengono utilizzati per prevedere i valori correnti della serie dipendente. Per esempio, un ordine numeratore 1 specifica che per prevedere il valore corrente di ogni serie dipendente viene utilizzato il valore di una serie indipendente di un periodo precedente, così come il valore corrente della serie indipendente.

Denominatore. L'ordine del denominatore della funzione di trasferimento indica come vengono utilizzate le deviazioni dalla media della serie per i valori precedenti della serie indipendente (predittore) selezionata per prevedere i valori correnti della serie dipendente. Per esempio, un ordine denominatore 1 specifica di considerare le deviazioni dal valore della media di una serie indipendente di un periodo precedente quando si prevede il valore corrente di ogni serie dipendente.

Differenza. Specifica l'ordine di differenziazione applicato alla serie indipendente (predittore) selezionata prima di eseguire la stima dei modelli. La differenziazione è necessaria quando sono presenti delle tendenze e viene utilizzata per rimuoverne l'effetto.

Gradi stagionali. I componenti numeratore, denominatore e differenziazione stagionali, hanno lo stesso ruolo delle corrispondenti controparti non stagionali. Per gli ordini stagionali tuttavia, i valori di serie correnti vengono influenzati dai valori di serie precedenti separati da uno o più periodi stagionali. Ad esempio, per i dati mensili (periodo stagionale di 12), un ordine stagionale 1 è il valore della serie corrente è influenzato dal valore della serie che precede di 12 periodi quello corrente. Specificare un ordine stagionale 1, per i dati mensili, è quindi come specificare un ordine non stagionale 12.

Ritardo. Questa impostazione determina un ritardo dell'influenza del campo di input pari al numero di intervalli specificato. Per esempio, se il ritardo è impostato su 5, il valore del campo di input al tempo t non influisce sulle previsioni fino a quando non sono trascorsi cinque periodi ($t + 5$).

Trasformazione. L'indicazione di una funzione di trasferimento per un insieme di variabili indipendenti comprende anche una trasformazione facoltativa da eseguire su tali variabili.

- **Nessuna.** Non viene eseguita nessuna trasformazione.
- **Radice quadrata.** Viene eseguita la trasformazione radice quadrata.

- **Log naturale.** Viene eseguita la trasformazione logaritmica naturale.

Gestione dei valori anomali

La scheda Valori anomali offre diverse opzioni per la gestione dei valori anomali nei dati ³.

Non individuare valori anomali o modellarli. Per impostazione predefinita, i valori anomali non vengono né rilevati né modellati. Selezionare questa opzione per disattivare qualsiasi rilevamento o modello di valori anomali.

Rileva automaticamente valori anomali. Selezionare questa opzione per rilevare automaticamente i valori anomali e scegliere uno o più tipi di valori anomali visualizzati.

Tipo di valori anomali da rilevare. Selezionare i tipi di valori anomali da rilevare. I tipi supportati sono:

- Additiva (default)
- Cambiamento di livello (default)
- Innovativo
- Transiente
- Additiva stagionale
- Tendenza locale
- Patch additiva

Per ulteriori informazioni consultare l'argomento *Valori anomali* nel documento *IBM SPSS Modeler 16 - Nodi di modellazione*.

Opzioni di distribuzione di Streaming TS

Se si utilizza il nodo Streaming TS in un ambiente di distribuzione in streaming, tramite IBM SPSS Modeler Solution Publisher, per default ogni volta che viene utilizzato un modello di serie temporale per calcolare il punteggio dei record, il modello viene creato nuovamente e vengono effettuate nuove previsioni. È possibile modificare questo comportamento di default utilizzando le opzioni della scheda Distribuzione per specificare quando creare nuovamente i modelli di serie temporali.

Nota: Le impostazioni della scheda Distribuzione vengono ignorate per il normale utilizzo del nodo Streaming TS in IBM SPSS Modeler.

Forza ricostruzione completa quando necessario (può cambiare tipo di modello). Selezionare questa opzione se si desidera che i modelli di serie temporali vengano creati di nuovo solo se è presente un numero specificato di nuovi record. Creare nuovamente il modello ogni terzo o quarto nuovo record piuttosto che ad ogni nuovo record può migliorare le prestazioni. La selezione di questa opzione consente i controlli **Numero minimo di record nuovi che saranno necessari per la ricostruzione**.

Numero minimo di record nuovi che saranno necessari per la ricostruzione. È possibile scegliere quando creare nuovamente i modelli di serie temporali in base al numero di nuovi record. Selezionare una delle seguenti opzioni:

- Se **Utilizza la percentuale di record (%)** è selezionato, i modelli di serie temporali vengono nuovamente creati quando una percentuale specificata del numero totale di record è costituita da nuovi record. La percentuale di default è 10%.
- Se è selezionato **Utilizza il numero assoluto di record**, il modello di serie temporale viene nuovamente creato quando un numero specificato di record è costituito da nuovi record. Il numero minimo di nuovi record che può essere specificato per la nuova creazione del modello è uno e il numero massimo è 100. Il numero di record di default è quattro.

3. Pena, D., G. C. Tiao, and R. S. Tsay, eds. 2001. *A course in time series analysis*. New York: John Wiley and Sons.

Campo per identificare nuovi record. Selezionare il campo che verrà osservato per rilevare quando vi sono nuovi record. Il campo scelto deve essere quello che ordina i record nella serie temporale. È possibile scegliere solo i campi continui.

Nodo STB (Space-Time-Box)

Gli STB (Space-Time-Box) sono un'estensione delle posizioni di spazio con Geohash. Più specificatamente, un STB è una stringa alfanumerica che rappresenta un'area delimitata regolarmente di spazio e tempo.

Ad esempio, l'STB: dr5ru7|2013-01-01 00:00:00|2013-01-01 00:15:00 è costituito dalle seguenti 3 parti:

- Il geohash - dr5ru7
- Il timestamp di inizio - 2013-01-01 00:00:00
- Il timestamp di fine - 2013-01-01 00:15:00

Come esempio, si potrebbero utilizzare le informazioni di spazio e tempo per migliorare la confidenza che due entità sono identiche perché sono virtualmente nello stesso posto nello stesso momento. In alternativa si potrebbe migliorare l'accuratezza dell'identificazione delle relazioni mostrando che due entità sono correlate per la loro vicinanza nello spazio e nel tempo.

È possibile scegliere la modalità **Record individuali** o **Hangout** secondo quanto appropriato per i requisiti. Entrambe le modalità richiedono gli stessi dettagli di base, riportati di seguito:

Campo latitudine. Selezionare il campo che identifica la latitudine.

Campo longitudine. Selezionare il campo che identifica la longitudine.

Campo timestamp. Selezionare il campo che identifica l'ora e la data.

Opzioni record individuale

Utilizzare questa modalità per aggiungere un ulteriore campo ad un record per identificare la sua posizione in un momento specifico.

Ricava. Selezionare una o più densità di spazio e tempo da cui ricavare il nuovo campo. Consultare "Definizione della densità STB (Space-Time-Box)" a pagina 93 per ulteriori informazioni.

Estensione nome campo. Digitare l'estensione da aggiungere ai nuovi nomi di campo. È possibile scegliere di aggiungere questa estensione come **Suffisso** o **Prefisso**.

Opzioni di hangout

Un hangout può essere visto come una posizione e/o un tempo in cui un'entità viene individuata continuamente o ripetutamente. Ad esempio, potrebbe essere utilizzato per identificare un veicolo che effettua un trasporto regolare ed identifica eventuali deviazioni dalla norma.

Densità STB. Selezionare la densità di spazio e tempo da cui ricavare il nuovo campo. Consultare "Definizione della densità STB (Space-Time-Box)" a pagina 93 per ulteriori informazioni.

Campo ID entità. Selezionare l'entità da utilizzare come identificativo hangout.

Numero minimo di eventi. Un evento è una riga nei dati. Selezionare il numero minimo di occorrenze di un evento per l'entità per poterla considerare in hangout.

Tempo di attesa minimo. Specificare il periodo minimo durante il quale l'entità deve rimanere nella stessa posizione. Questo è utile per evitare, ad esempio, di considerare in hangout una macchina ferma ad un semaforo.

Consenti hangout oltre i limiti STB. Se questa opzione è selezionata, la definizione di un hangout è meno rigida e potrebbe includere, ad esempio, un'entità che si sofferma in più di un STB (Space-Time-Box). Ad esempio, se gli STB sono definiti come ore intere, selezionando questa opzione verrebbe considerata come valida un'entità che è in hangout per un'ora, anche se l'ora è costituita dai trenta minuti precedenti alla mezzanotte e i trenta minuti successivi alla mezzanotte. Se questa opzione non è selezionata, il 100% del tempo di hangout deve essere compreso in un singolo Space-Time-Box.

Proporzione minima di eventi in timebox valido (%). Disponibile solo se è selezionato **Consenti hangout oltre i limiti STB**. Utilizzarlo per controllare il grado rispetto al quale un hangout segnalato in un STB potrebbe di fatto sovrapporsi ad un altro. Selezionare la proporzione minima di eventi che devono verificarsi in un singolo STB per identificare un hangout. Ad esempio, se è impostato su 25%, e la proporzione degli eventi è 26%, viene qualificato come hangout.

Definizione della densità STB (Space-Time-Box)

Scegliere la dimensione (densità) dell'STB (Space-Time-Box) specificando l'area fisica e il tempo trascorso da includere in ogni STB.

Densità Geo. Selezionare la dimensione dell'area da includere in ogni STB.

Intervallo di tempo. Selezionare il numero di ore da includere in ogni STB.

Nome campo. Con prefisso STB, viene completato automaticamente in base alle selezioni nei due campi precedenti.

Capitolo 4. Nodi Operazioni su campi

Panoramica sui nodi Operazioni sui campi

Dopo un'esplorazione dati iniziale, probabilmente sarà necessario selezionare, ripulire o creare i dati per predisporli all'analisi. La palette Operazioni su campi contiene numerosi nodi utili per le operazioni di trasformazione e preparazione.

Per esempio, utilizzando un nodo Ricava, è possibile creare un attributo che attualmente non è rappresentato nei dati. In alternativa, è possibile utilizzare un nodo Discretizza per ricodificare automaticamente i valori di campo per le analisi selezionate. Il nodo Tipo è uno dei nodi utilizzati più frequentemente, perché consente di assegnare un livello di misurazione, i valori ed un ruolo di modellazione per ciascun campo nell'insieme di dati. Queste operazioni sono utili per gestire i valori mancanti e la creazione di modelli a valle.

La palette Operazioni su campi contiene i seguenti nodi:



Il nodo Preparazione automatica dati (ADP) può analizzare i dati e individuare le correzioni, escludere i campi problematici o probabilmente inutili e derivare all'occorrenza nuovi attributi, migliorando le performance grazie allo screening intelligente e alle tecniche di campionamento. Il nodo si può utilizzare in modo completamente automatico, permettendogli di scegliere e di applicare le correzioni, oppure visualizzando in anteprima le modifiche prima dell'applicazione e accettandole, respingendole o modificandole a seconda dei casi.



Il nodo Tipo specifica proprietà e metadati di campo. Per esempio, è possibile specificare un livello di misurazione (continuo, nominale, ordinale o flag) per ogni campo, impostare opzioni relative alla gestione dei valori mancanti e dei valori null di sistema, impostare il ruolo di un campo per la modellazione, specificare le etichette di campo e valore e specificare i valori per un campo.



Il nodo Filtro filtra (ignora) campi, rinomina campi e mappa campi tra i nodi origine.



Il nodo Ricava modifica valori di dati o crea nuovi campi da uno o più campi esistenti. Crea campi di tipo Formula, Flag, Nominale, Stato, Conteggio e Condizionale.



Il nodo dell'insieme combina due o più nugget del modello al fine di ottenere previsioni più precise di quelle ricavabili dai singoli modelli.



Il nodo Riempimento sostituisce valori di campo e modifica l'archiviazione. È possibile scegliere di sostituire i valori in base a una condizione CLEM, per esempio @BLANK(@FIELD). In alternativa, si può scegliere di sostituire tutti i valori null o vuoti con un valore specifico. Il nodo Riempimento è utilizzato spesso in combinazione con il nodo Tipo per sostituire valori mancanti.



Il nodo Anonimizza consente di mascherare i nomi o i valori dei campi, quando si utilizzano dati da includere in un modello a valle del nodo, permettendo di nascondere i dati originali. Questa funzionalità può essere utile se si desidera consentire ad altri utenti di creare modelli utilizzando dati riservati, quali nomi di clienti o altri dettagli.



Il nodo Ricodifica trasforma un insieme di valori categoriali in un altro. L'operazione di ricodifica consente di comprimere categorie o raggruppare dati per l'analisi.



Il nodo Discretizza crea automaticamente nuovi campi nominali (insieme) basati sui valori di uno o più campi continui (intervallo numerico) esistenti. Per esempio, è possibile trasformare un campo continuo relativo al reddito in campo categoriale contenente gruppi di reddito come deviazioni dalla media. Dopo aver creato bin per il nuovo campo, è possibile generare un nodo Ricava basato sui punti di divisione.



Il nodo Analisi RFM (Recency, Frequency, Monetary, Passato recente, Frequenza, Monetario) consente di determinare in modo quantitativo i clienti potenzialmente migliori verificando quanto tempo è trascorso dal loro ultimo acquisto (passato recente), con quale frequenza hanno effettuato acquisti (frequenza) e quanto hanno speso per tutte le transazioni (monetario).



Il nodo Partizione genera un campo partizione che suddivide i dati in sottoinsiemi separati per le fasi di addestramento, verifica e convalida della creazione del modello.



Il nodo Crea flag crea campi flag in base ai valori categoriali di uno o più campi nominali.



Il nodo Riorganizza converte un campo nominale o flag in un gruppo di campi in cui è possibile inserire i valori di un altro campo. Per esempio, dato un campo denominato *tipo di pagamento*, con valori di *credito*, *contanti* e *debito*, verrebbero creati tre nuovi campi (*credito*, *contanti*, *debito*), ognuno dei quali può contenere il valore del pagamento effettuato.



Il nodo Trasponi scambia i dati delle righe e delle colonne in modo da trasporre i campi in record e i record in campi.



Il nodo Intervalli di tempo specifica intervalli e, se necessario, crea etichette per la modellazione di dati di serie temporali. Se i valori non sono spazati in modo uniforme, il nodo può riempire o aggregare i valori in base alle proprie esigenze per generare un intervallo uniforme tra i record.



Il nodo Cronologia crea nuovi campi contenenti dati dei campi di record precedenti. I nodi Cronologia sono utilizzati in genere per dati sequenziali, per esempio per dati di serie temporali. Prima di utilizzare un nodo Cronologia, può essere utile ordinare i dati con un nodo Ordina.



Il nodo Ordina campi definisce l'ordine naturale utilizzato per visualizzare i campi a valle. Tale ordine incide sulla visualizzazione dei campi in vari contesti, quali tabelle, elenchi e Selettore di campo. Questa operazione risulta utile se si desidera rendere più visibili i campi interessanti in insiemi di dati di grandi dimensioni.

Molti di questi nodi possono essere generati direttamente dal report di verifica creato da un nodo Verifica dati. Per ulteriori informazioni, consultare l'argomento "Generazione di altri nodi per la preparazione dei dati" a pagina 290.

Preparazione automatica dati

La preparazione dei dati per l'analisi rappresenta una delle fasi più importanti in qualsiasi progetto — e, tradizionalmente, una delle attività che richiedono più tempo. La funzione Preparazione automatica dati (ADP) svolge questo compito al posto dell'utente, analizzando i dati e individuando le correzioni da apportare, escludendo i campi problematici o probabilmente inutili, derivando nuovi attributi se necessario e migliorando le prestazioni attraverso tecniche di screening intelligenti. È possibile utilizzare l'algoritmo in modo completamente **automatico**, consentendo all'algoritmo di scegliere ed applicare le correzioni oppure è possibile utilizzare la modalità **interattiva**, in cui viene visualizzata un'anteprima delle modifiche prima che vengano apportate, in modo che sia possibile accettarle o rifiutarle in base alle proprie esigenze.

L'utilizzo di ADP consente di predisporre i dati per la creazione del modello in modo semplice e rapido, senza che sia necessario conoscere i concetti statistici impiegati. La creazione e il calcolo del punteggio dei modelli tenderanno a essere più rapidi; inoltre, l'utilizzo di ADP migliora la robustezza dei processi di modellazione automatica quali l'aggiornamento dei modelli e il metodo campione / sfidante.

Nota: quando ADP prepara un campo per l'analisi, crea un nuovo campo contenente le regolazioni o le trasformazioni, invece di sostituire i valori e le proprietà esistenti dal campo precedente. Il campo precedente non viene utilizzato nelle analisi successive; il relativo ruolo è impostato su Nessuno.

Esempio. Una compagnia di assicurazioni con poche risorse per indagare sulle richieste di indennizzo dei proprietari immobiliari vuole creare un modello per evidenziare le richieste sospette e potenzialmente fraudolente. Prima di procedere, viene effettuata la preparazione automatica dei dati per la creazione del modello. Dal momento che la compagnia ha necessità di esaminare le trasformazioni proposte prima che queste vengano applicate, utilizzerà la preparazione automatica dati in modalità interattiva.

Un gruppo industriale automobilistico tiene traccia delle vendite per un'ampia gamma di autoveicoli personali. Nel tentativo di identificare modelli a basso e alto rendimento è possibile stabilire una relazione tra la vendita dei veicoli e le rispettive caratteristiche. Verrà utilizzata la preparazione automatica dei dati per l'analisi e verranno creati modelli utilizzando i dati "prima" e "dopo" la preparazione per scoprire come cambiano i risultati.

Qual è il proprio obiettivo? La Preparazione automatica dati consiglia una serie di fasi di preparazione dei dati che influiscono sulla velocità con cui altri algoritmi creano modelli e ne migliorano il potere predittivo. Possono comprendere funzioni di trasformazione, creazione e selezione. Anche l'obiettivo può essere trasformato. È possibile specificare le priorità di creazione dei modelli su cui deve concentrarsi il processo di preparazione dei dati.

- **Bilancia velocità e accuratezza.** Questa opzione prepara i dati in modo da dare la stessa priorità alla velocità di elaborazione dei dati da parte degli algoritmi di creazione del modello e alla precisione delle previsioni.
- **Ottimizza per velocità.** Questa opzione prepara i dati in modo da dare la priorità alla velocità di elaborazione dei dati da parte degli algoritmi di creazione del modello. Selezionare questa opzione quando si utilizzano insiemi di dati molto grandi o quando si desidera ottenere una risposta rapida.

- **Ottimizza per precisione.** Questa opzione prepara i dati in modo da dare la priorità alla precisione delle previsioni generate dagli algoritmi di creazione del modello.
- **Analisi personalizzata.** Selezionare questa opzione se si desidera modificare manualmente l'algoritmo nella scheda Impostazioni. Si noti che questa impostazione viene selezionata automaticamente se in seguito si apportano modifiche incompatibili con uno degli altri obiettivi alle opzioni della scheda Impostazioni.

Training del nodo

Il nodo ADP viene implementato come nodo di elaborazione e funziona in modo simile al nodo Tipo; **effettuare il training** del nodo ADP equivale a istanziare il nodo Tipo. Una volta eseguita l'analisi, le trasformazioni specificate vengono applicate ai dati senza ulteriore analisi, purché il modello di dati a monte resti invariato. Analogamente ai nodi Tipo e Filtro, alla disconnessione il nodo ADP ricorda il modello dati e le trasformazioni, in modo che alla successiva connessione non debba più essere sottoposto a training; in questo modo è possibile eseguirne il training su un sottoinsieme di dati tipici e quindi copiarlo o distribuirlo per l'uso su dati in tempo reale per il numero di volte necessario.

Utilizzo della barra degli strumenti

La barra degli strumenti consente di eseguire e aggiornare la visualizzazione dell'analisi dei dati e generare nodi utilizzabili congiuntamente ai dati originali.

- **Genera** Da questo menu è possibile generare un nodo Filtro o Ricava. Si noti che questo menu è disponibile solo quando nella scheda Analisi è visualizzata un'analisi.
Il nodo Filtro rimuove i campi di input trasformati. Se si configura il nodo ADP in modo da lasciare i campi di input originali nell'insieme di dati, l'insieme di input originale verrà ripristinato consentendo l'interpretazione del campo punteggio in termini di input. Questo potrebbe essere utile, per esempio, se si desidera produrre un grafico del campo punteggio rispetto a vari input.
Il nodo Ricava può ripristinare l'insieme di dati originale e le unità obiettivo. È possibile generare un nodo Ricava solo quando il nodo ADP contiene un'analisi che ridimensiona un obiettivo intervallo (è selezionato il ridimensionamento Box-Cox nel pannello Prepara input e obiettivo). Non è possibile generare un nodo Ricava se l'obiettivo non è un intervallo o se non è selezionata la modifica della scala di Box-Cox. Per ulteriori informazioni, consultare l'argomento "Generazione di un nodo Ricava" a pagina 110.
- **Visualizza** Contiene opzioni che determinano quali elementi vengono visualizzati nella scheda Analisi. Tali opzioni includono i controlli per la modifica dei grafici e le selezioni di visualizzazione sia per il riquadro principale sia per le visualizzazioni collegate.
- **Anteprima** Visualizza un esempio delle trasformazioni che saranno applicate ai dati di input.
- **Analizza dati** Avvia un'analisi secondo le impostazioni correnti e visualizza i risultati nella scheda Analisi.
- **Cancella analisi** Elimina l'analisi esistente (disponibile solo quando esiste un'analisi corrente).

Stato del nodo

Lo stato del nodo ADP nell'area di IBM SPSS Modeler è indicato da una freccia o da un segno di spunta sull'icona che mostra se l'analisi è stata effettuata oppure no.

Scheda Campi

Per poter generare un modello, è necessario prima specificare i campi da utilizzare come obiettivi e come input. Con alcune eccezioni, tutti i campi Modelli utilizzano le informazioni sui campi di un nodo Tipo a monte. Se si utilizza un nodo Tipo per selezionare i campi obiettivo e di input, non è necessario cambiare nessuna delle impostazioni presenti in questa scheda.

Utilizza impostazioni nodo Tipo. Questa opzione indica al nodo di utilizzare le informazioni sui campi da un nodo Tipo a monte. Questa è l'opzione di default.

Utilizza impostazioni personalizzate. Questa opzione indica al nodo di utilizzare le informazioni sui campi specificate qui al posto di quelle date in un qualsiasi nodo Tipo a monte. Dopo avere selezionato questa opzione, specificare i campi nell'area sottostante come richiesto.

Obiettivo. Per i modelli che richiedono uno o più campi obiettivo, selezionare il campo o i campi obiettivo. Questa operazione è simile all'impostazione del ruolo di un campo su *Obiettivo* in un nodo Tipo.

Input. Selezionare il campo o i campi di input. Questa operazione è simile all'impostazione del ruolo di un campo su *Input* in un nodo Tipo.

Scheda Impostazioni

La scheda Impostazioni contiene vari gruppi di impostazioni che è possibile modificare per perfezionare l'elaborazione dei dati da parte dell'algoritmo. Se si apportano modifiche alle impostazioni predefinite che risultano incompatibili con gli altri obiettivi, la scheda Obiettivo viene aggiornata automaticamente per selezionare l'opzione **Personalizza analisi**.

Impostazioni campo

Utilizza campo frequenza. Questa opzione consente di selezionare un campo come ponderazione della frequenza. Utilizzarla se i record dei dati di training rappresentano più di un'unità ciascuno; ad esempio, se si utilizzano dei dati aggregati. I valori del campo dovrebbero essere rappresentati dal numero di unità rappresentate da ciascun record.

Utilizza campo peso. Questa opzione consente di selezionare un campo come ponderazione del caso. I pesi dei casi si utilizzano per tenere conto delle differenze nella varianza tra i vari livelli del campo di output.

Come gestire i campi esclusi dai modelli. Specifica cosa accade ai campi esclusi; è possibile scegliere di filtrarli per escluderli dai dati o semplicemente di impostarne il *ruolo* su **Nessuno**.

Nota: questa azione viene applicata anche all'obiettivo, se trasformato. Per esempio, se la nuova versione derivata dell'obiettivo è utilizzata come campo **Obiettivo**, l'obiettivo originale viene filtrato oppure impostato su **Nessuno**.

Se i campi in entrata non corrispondono all'analisi esistente. Specificare cosa accade se uno o più campi di input richiesti mancano dall'insieme di dati in arrivo quando si esegue un nodo ADP sottoposto a training.

- **Interrompi l'esecuzione e tieni l'analisi esistente.** Interrompe il processo di esecuzione, preserva le informazioni dell'analisi corrente e visualizza un errore.
- **Cancella l'analisi esistente e analizza i nuovi dati.** Cancella l'analisi esistente, analizza i dati in arrivo e applica le trasformazioni consigliate a tali dati.

Prepara date e ore

Molti algoritmi di modellazione non sono in grado di gestire direttamente i dettagli relativi a date e ore; queste impostazioni consentono di derivare nuovi dati sulle durate utilizzabili come input per i modelli dalle date e dalle ore indicate nei dati esistenti. I campi contenenti date e ore devono essere predefiniti con tipi di archiviazione data o ora. L'uso dei campi data e ora originali come input per i modelli in seguito alla preparazione automatica dei dati non è consigliato.

Prepara date e ore per la modellazione. Se si diseleziona questa opzione vengono disabilitati tutti gli altri controlli Prepara date e ore ma vengono mantenute le selezioni.

Calcola tempo trascorso fino alla data di riferimento. Produce il numero di anni/mesi/giorni a partire da una data di riferimento per ciascuna variabile che contiene delle date.

- **Data di riferimento.** Specifica la data a partire da cui sarà calcolata la durata relativamente alle informazioni sulla data presenti nei dati di input. Selezionando **Data odierna**, viene sempre utilizzata la data di sistema corrente quando viene eseguito ADP. Per utilizzare una data specifica, selezionare **Data fissa** e immettere la data desiderata. La data corrente viene immessa automaticamente nel campo **Data fissa** quando il nodo viene creato per la prima volta.
- **Unità per la durata della data.** Specificare se ADP deve decidere automaticamente l'unità per la durata della data oppure selezionarne una da **Unità fisse** in Anni, mesi o Giorni.

Calcola tempo trascorso fino all'ora di riferimento. Produce il numero di ore/minuti/secondi a partire da un'ora di riferimento per ciascuna variabile che contiene delle ore.

- **Ora di riferimento.** Specifica l'ora a partire dalla quale sarà calcolata la durata relativamente alle informazioni sull'ora presenti nei dati di input. Se si seleziona **Ora corrente**, viene sempre utilizzata l'ora corrente del sistema per l'esecuzione di ADP. Per utilizzare un'ora specifica, selezionare **Ora fissa** e immettere l'ora desiderata. L'ora corrente viene immessa automaticamente nel campo **Ora fissa** quando il nodo viene creato per la prima volta.
- **Unità per la durata dell'ora.** Specificare se ADP deve decidere automaticamente l'unità per la durata dell'ora oppure selezionarne una da **Unità fisse** in Ore, minuti o Secondi.

Estrai elementi di tempo ciclico. Utilizzare queste impostazioni per suddividere un singolo campo data o ora in uno o più campi. Ad esempio, se vengono selezionate tutte e tre le caselle di controllo della data, il campo della data di input "1954-05-23" viene suddiviso in tre campi: 1954, 5 e 23, ciascuno dei quali utilizza il suffisso definito nel pannello **Nomi di campi** ed il campo della data originale viene ignorato.

- **Estrai dalle date.** Per qualsiasi input di data, specificare se si desidera estrarre gli anni, i mesi, i giorni o una combinazione dei tre elementi.
- **Estrai dalle ore.** Per qualsiasi input di ora, specificare se si desidera estrarre le ore, i minuti, i secondi o una combinazione dei tre elementi.

Escludi campi

La scarsa qualità dei dati può influire sulla precisione delle previsioni; pertanto, è possibile specificare il livello di qualità accettabile per le funzioni di input. Tutti i campi che sono costanti o hanno il 100% dei valori mancanti vengono esclusi automaticamente.

Escludi campi di input a bassa qualità. Se si diseleziona questa opzione vengono disabilitati tutti gli altri controlli Escludi campi ma vengono mantenute le selezioni.

Escludi campi con troppi valori mancanti. I campi con una percentuale di valori mancanti superiore a quella specificata vengono eliminati dalla successiva analisi. Specificare un valore superiore o uguale a 0, che equivale a diselezionare questa opzione, e inferiore o uguale a 100, benché i campi con tutti i valori mancanti vengano automaticamente esclusi. Il valore di default è 50.

Escludi campi nominali con troppe categorie univoche. I campi nominali con un numero di categorie superiore a quello specificato vengono eliminati dalla successiva analisi. Specificare un intero positivo. Il valore di default è 100. È utile per rimuovere automaticamente i campi che contengono informazioni univoche del record provenienti dalla modellazione, come ID, indirizzo o nome.

Escludi campi categoriali con troppi valori in una categoria singola. I campi ordinali e nominali con una categoria contenente una percentuale di record superiore a quella specificata vengono eliminati dalla

successiva analisi. Specificare un valore superiore o uguale a 0, che equivale a deselezionare questa opzione, e inferiore o uguale a 100, benché i campi costanti vengano automaticamente esclusi. Il valore di default è 95.

Preparazione degli input e degli obiettivi

Dal momento che nessun dato si trova mai in uno stato perfetto per l'elaborazione, è possibile regolare alcune delle impostazioni prima di eseguire un'analisi. Ad esempio, è possibile rimuovere i valori anomali, specificando come devono essere gestiti i valori mancanti o correggendo il tipo.

Nota: se i valori in questo pannello vengono modificati, la scheda **Obiettivi** viene aggiornata automaticamente per selezionare l'opzione **Analisi personalizzata**.

Prepara i campi di input e obiettivo per la modellazione. Attiva/disattiva tutti i campi nel riquadro.

Modifica tipo e migliora qualità dati. Per gli input e l'obiettivo è possibile specificare separatamente diverse trasformazioni dati; ciò è utile quando non si desidera modificare i valori dell'obiettivo. Ad esempio, una previsione di reddito in dollari ha più senso di una misurata in log(dollari). Inoltre, se l'obiettivo ha dei valori mancanti, non è presente alcun guadagno previsto per riempire i valori mancanti, mentre riempire i valori mancanti negli input può rendere possibili alcuni algoritmi di elaborazione delle informazioni che altrimenti andrebbero persi.

Altre impostazioni per queste trasformazioni, quale ad esempio il valore di interruzione anomalo, sono comuni sia all'obiettivo che agli input.

È possibile selezionare le impostazioni per gli input, l'obiettivo o entrambe le categorie:

- **Modifica il tipo di campi numerici.** Selezionare questa opzione per determinare se i campi numerici con il livello di misurazione *Ordinale* possono essere convertiti in *Continuo*, o viceversa. Per controllare la conversione, si specificano i valori di soglia minimo e massimo.
- **Riordina campi nominali.** Selezionare questa opzione per ordinare i campi (insieme) nominali, dalla categoria minima a quella massima.
- **Sostituisci valori anomali nei campi continui.** Specifica se sostituire i valori anomali; utilizzare questa opzione in combinazione con le opzioni **Metodo di sostituzione dei valori anomali** seguenti.
- **Campi continui: sostituisci i valori mancanti con media.** Selezionare questa opzione per sostituire i valori mancanti delle funzioni (intervallo) continue.
- **Campi nominali: sostituisci i valori mancanti con modalità** Selezionare questa opzione per sostituire i valori mancanti delle funzioni (insieme) nominali.
- **Campi ordinali: sostituisci i valori mancanti con mediana** Selezionare questa opzione per sostituire i valori mancanti delle funzioni (insieme ordinato) ordinali.

Numero massimo di valori per i campi ordinali. Specifica la soglia per ridefinire i campi (insieme ordinato) ordinali come continui (intervallo). Il valore predefinito è 10; quindi, se un campo ordinale ha più di 10 categorie viene ridefinito come continuo (intervallo).

Numero minimo di valori per i campi continui. Specifica la soglia per ridefinire i campi scala o continui (intervallo) come ordinali (insieme ordinato). Il valore predefinito è 5; quindi, se un campo continuo ha meno di 5 valori viene ridefinito come ordinale (insieme ordinato).

Valore di interruzione anomalo. Specifica il criterio di interruzione dei valori anomali, misurato in deviazioni standard; il valore predefinito è 3.

Metodo di sostituzione dei valori anomali. Selezionare se i valori anomali devono essere sostituiti ritagliandoli (coercizione) con il punto di interruzione o eliminati impostandoli come valori mancanti. Tutti i valori anomali impostati su valori mancanti rispondono alle impostazioni per la gestione dei valori mancanti selezionate in precedenza.

Mettili tutti i campi di input continui in una scala comune. Per normalizzare i campi di input continui, selezionare questa casella di controllo e scegliere il metodo di normalizzazione. Il valore predefinito è **trasformazione punteggio Z**, dove è possibile specificare un valore per **Media finale**, il cui valore predefinito è 0, e per **Deviazione standard finale**, il cui valore predefinito è 1. In alternativa, è possibile scegliere di utilizzare **Trasformazione Min/Max** e specificare i valori minimo e massimo, i cui valori predefiniti sono rispettivamente 0 e 100.

Questo campo è particolarmente utile quando si seleziona **Esegui creazione funzioni** nel pannello Crea e seleziona funzioni.

Ridimensiona target continuo con trasformazione di Box-Cox. Per normalizzare un campo obiettivo (scala o intervallo) continuo, selezionare questa casella di controllo. I valori predefiniti della trasformazione di Box-Cox sono 0 per **Media finale** e 1 per **Deviazione standard finale**.

Nota: se si decide di normalizzare l'obiettivo, la dimensione dell'obiettivo verrà trasformata. In questo caso può essere necessario generare un nodo Ricava per applicare una trasformazione inversa al fine di riconvertire le unità trasformate in un formato riconoscibile a scopi di ulteriori elaborazioni. Per ulteriori informazioni, consultare l'argomento "Generazione di un nodo Ricava" a pagina 110.

Creazione e selezione delle funzioni

Per migliorare il potere predittivo dei dati è possibile trasformare i campi di input o crearne di nuovi basati su quelli esistenti.

Nota: se i valori in questo pannello vengono modificati, la scheda **Obiettivi** viene aggiornata automaticamente per selezionare l'opzione **Analisi personalizzata**.

Trasformare, creare e selezionare campi di input per rafforzare il potere predittivo. Attiva/disattiva tutti i campi nel riquadro.

Unisci categorie sparse per aumentare al massimo l'associazione all'obiettivo. Selezionare questa opzione per creare un modello più gestibile riducendo il numero delle variabili da elaborare in associazione all'obiettivo. Se richiesto, modificare il valore predefinito 0,05 della probabilità.

Se tutte le categorie vengono unite in una, le versioni originale e derivata del campo vengono escluse perché non hanno un valore come predittore.

Quando non ci sono obiettivi, unisci le categorie sparse in base ai conteggi. Se si utilizzano dati privi di obiettivo, è possibile scegliere di unire categorie sparse di funzioni ordinali (insieme ordinato) e/o nominali (insieme). Specificare la percentuale minima di casi, o record, nei dati che identificano le categorie da unire; il valore predefinito è 10.

Le categorie vengono unite mediante le seguenti regole:

- L'unione non viene effettuata sui campi binari.
- Se durante l'unione sono presenti solo due categorie, l'unione si interrompe.
- Se non è presente alcuna categoria originale né alcuna categoria creata durante l'unione, con una percentuale di casi inferiori a quella minima specificata, l'unione si interrompe.

Categorizza i campi continui e mantieni il potere predittivo. Se l'insieme di dati comprende un obiettivo categoriale, è possibile categorizzare gli input continui con associazioni forti per migliorare le prestazioni in sede di elaborazione. Se richiesto, modificare il valore predefinito 0,05 della probabilità per i sottoinsiemi omogenei.

Se con l'operazione di categorizzazione si ottiene un unico intervallo per un determinato campo, la versione originale e categorizzata del campo vengono escluse perché sono prive di valore come predittori.

Nota: la raccolta in ADP è differente dalla raccolta ottimale utilizzata in altre parti di IBM SPSS Modeler. La categorizzazione ottimale utilizza le informazioni relative all'entropia per convertire una variabile continua in una categoriale; per far questo è necessario ordinare i dati e archivarli tutti in memoria. ADP utilizza sottoinsiemi omogenei per categorizzare una variabile continua e, pertanto, la categorizzazione in ADP non richiede di ordinare i dati e non li archivia tutti in memoria. Quando si utilizza il metodo dei sottoinsiemi omogenei per categorizzare una variabile continua, il numero di categorie dopo la categorizzazione è sempre inferiore o uguale al numero di categorie dell'obiettivo.

Effettua selezione delle funzioni. Selezionare questa opzione per rimuovere le funzioni con un coefficiente di correlazione basso. Se richiesto, modificare il valore predefinito 0,05 della probabilità.

Questa opzione viene applicata solo alle funzioni di input continue in cui l'obiettivo è continuo e alle funzioni di input categoriali.

Esegui creazione funzioni. Selezionare questa opzione per derivare nuove funzioni da una combinazione di varie funzioni esistenti (che vengono quindi escluse dalla modellazione).

Questa opzione viene applicata solo alle funzioni di input continue quando l'obiettivo è continuo oppure non esiste.

Nomi di campi

Per individuare facilmente le funzioni nuove e trasformate, ADP crea e applica nuovi nomi, prefissi o suffissi di base. I nomi si possono modificare in modo da renderli più pertinenti rispetto alle esigenze e ai dati dell'utente. Se si desidera specificare altre etichette, è necessario effettuare tale operazione in un nodo Tipo a valle.

Campi trasformati e creati. Specificare le estensioni dei nomi da applicare ai campi obiettivo e di input trasformati.

Si noti che nel nodo ADP l'impostazione dei campi stringa come vuoti può provocare un errore, a seconda di come si decide di gestire i campi non utilizzati. Se **Come gestire i campi esclusi dai modelli** è impostato su **Filtra i campi inutilizzati** nel riquadro Impostazioni campo della scheda Impostazioni, le estensioni dei nomi per gli input e l'obiettivo si possono impostare su "nothing". I campi originali vengono esclusi e i campi trasformati vengono salvati al loro posto; in questo caso i nuovi campi trasformati avranno lo stesso nome degli originali.

Tuttavia, se si seleziona **Imposta direzione campi inutilizzati su 'Nessuna'**, le estensioni nome vuote o null per l'obiettivo e gli input causano un errore perché in questo modo viene eseguito un tentativo di creare nomi di campi duplicati.

Specificare inoltre il nome del prefisso da applicare a tutte le funzioni da creare con le impostazioni Seleziona e Crea. Il nuovo nome viene creato apponendo un suffisso numerico al nome radice del prefisso. Il formato del numero dipende dal numero di nuove funzioni da derivare, ad esempio:

- le funzioni create da 1 a 9 verranno denominate: da feature1 a feature9.
- le funzioni create da 10 a 99 verranno denominate: da feature01 a feature99.
- le funzioni create da 100 a 999 verranno denominate: da feature001 a feature999 e così via.

In questo modo, le funzioni create saranno organizzate in un ordine logico indipendentemente dal loro numero.

Durate calcolate da date e ore. Specificare le estensioni dei nomi da applicare alle durate calcolate a partire da date e ore.

Elementi ciclici estratti da date e ore. Specificare le estensioni dei nomi da applicare agli elementi ciclici estratti da date e ore.

Scheda Analisi

1. Quando le impostazioni di ADP sono soddisfacenti (comprese le eventuali modifiche apportate alla scheda Obiettivo, Campi e Impostazioni), fare clic su **Analizza dati**; l'algoritmo applica le impostazioni ai dati immessi e visualizza i risultati nella scheda Analisi.

La scheda Analisi contiene output in formato tabellare e grafico che riassume l'elaborazione dei dati e visualizza le raccomandazioni su eventuali modifiche o miglioramenti da apportare ai dati per il calcolo del punteggio. Questo consente di esaminare e accettare o rifiutare tali raccomandazioni.

La scheda Analisi è composta da due riquadri, la visualizzazione principale a sinistra e quella collegata o ausiliaria a destra. Le visualizzazioni principali sono tre:

- Riepilogo elaborazione campi (impostazione predefinita). Per ulteriori informazioni, consultare l'argomento "Riepilogo elaborazione campi".
- Campi. Per ulteriori informazioni, consultare l'argomento "Campi" a pagina 105.
- Riepilogo delle azioni. Per ulteriori informazioni, consultare l'argomento "Riepilogo delle azioni" a pagina 106.

Le visualizzazioni collegate/ausiliarie sono quattro:

- Potere predittivo (impostazione predefinita). Per ulteriori informazioni, consultare l'argomento "Potere predittivo" a pagina 106.
- Tabella campi. Per ulteriori informazioni, consultare l'argomento "Tabella Campi" a pagina 106.
- Dettagli campo. Per ulteriori informazioni, consultare l'argomento "Dettagli campo" a pagina 107.
- Dettagli dell'azione. Per ulteriori informazioni, consultare l'argomento "Dettagli azione" a pagina 108.

Collegamenti tra le visualizzazioni

All'interno della visualizzazione principale, il testo sottolineato nelle tabelle controlla la visualizzazione nella visualizzazione collegata. Se si fa clic sul testo è possibile visualizzare i dettagli di un determinato campo, insieme di campi o fase di elaborazione. Il collegamento selezionato per ultimo è visualizzato con un colore più scuro per facilitare l'individuazione del collegamento tra il contenuto dei due riquadri.

Reimpostazione delle visualizzazioni

Per visualizzare nuovamente le raccomandazioni originali della scheda Analisi e annullare le eventuali modifiche apportate alle visualizzazioni Analisi, fare clic su **Reimposta** nella parte inferiore del riquadro di visualizzazione principale.

Riepilogo elaborazione campi

La tabella Riepilogo elaborazione campi offre una snapshot dell'impatto complessivo stimato dell'elaborazione, comprese le modifiche dello stato delle funzioni e il numero di funzioni create.

Si noti che non viene effettivamente creato alcun modello e, pertanto, non vi è alcuna misura o grafico della variazione del potere predittivo generale prima e dopo la preparazione dei dati; è possibile invece visualizzare i grafici del potere predittivo dei singoli predittori consigliati.

La tabella riporta le seguenti informazioni:

- Il numero di campi obiettivo.
- Il numero di predittori (input) originali.
- I predittori consigliati per l'uso nell'analisi e nella modellazione. Sono inclusi il numero totale di campi consigliati; il numero di campi originali, non trasformati consigliati; il numero di campi trasformati consigliati (escluse le versioni intermedie di qualsiasi campo, i campi derivati dai predittori di data/ora e i predittori creati); il numero di campi consigliati derivati dai campi data/ora e il numero di predittori creati consigliati.

- Il numero di predittori di input non consigliato per l'uso in nessuna forma, che si tratti della forma originale, come campo derivato o come input per un predittore creato.

Se le informazioni dei **Campi** sono sottolineate, farvi clic sopra per visualizzare ulteriori dettagli in una visualizzazione collegata. I dettagli relativi a **Obiettivo**, **Funzioni di input** e **Funzioni di input non utilizzate** sono riportati nella visualizzazione collegata Tabella campi. Per ulteriori informazioni, consultare l'argomento "Tabella Campi" a pagina 106. **Le funzioni consigliate per l'utilizzo nell'analisi** sono visualizzate nella visualizzazione collegata Potere predittivo. Per ulteriori informazioni, consultare l'argomento "Potere predittivo" a pagina 106.

Campi

La visualizzazione principale Campi mostra i campi elaborati e indica se ADP ne consiglia o meno l'utilizzo nei modelli a valle. È possibile ignorare le raccomandazioni di tutti i campi, ad esempio per escludere funzioni create o includere funzioni di cui ADP consiglia l'esclusione. Se un campo è stato trasformato, è possibile decidere se accettare la trasformazione suggerita o utilizzare la versione originale.

La visualizzazione Campi è composta da due tabelle, una per l'obiettivo e una per i predittori elaborati o creati.

Tabella Obiettivo

La tabella **Obiettivo** è visualizzata solo se nei dati è stato definito un obiettivo.

La tabella contiene due colonne:

- **Nome.** Si tratta del nome o dell'etichetta del campo obiettivo. Viene sempre utilizzato il nome originale, anche se il campo è stato trasformato.
- **Livello di misurazione.** In questa colonna è visualizzata l'icona che rappresenta il livello di misurazione; passare il puntatore del mouse sopra l'icona per visualizzare un'etichetta (continuo, ordinale, nominale e così via) che descrive i dati.

Se l'obiettivo è stato trasformato, la colonna **Livello di misurazione** riflette la versione trasformata finale. *Nota:* non è possibile disattivare le trasformazioni per l'obiettivo.

Tabella Predittori

La tabella **Predittori** è sempre visualizzata. Ogni riga della tabella rappresenta un campo. Per impostazione predefinita, le righe sono ordinate in modo decrescente in base al potere predittivo.

Per le funzioni ordinarie, il nome originale viene sempre utilizzato come nome della riga. Nella tabella sono riportate sia le versioni originali che quelle derivate dei campi data/ora (in righe separate); la tabella comprende anche i predittori creati.

Si noti che le versioni trasformate dei campi visualizzate nella tabella rappresentano sempre le versioni finali.

Per impostazione predefinita, nella tabella Predittori sono visualizzati solo i campi consigliati. Per visualizzare gli altri campi, selezionare la casella **Includi campi non raccomandati nella tabella** sopra la tabella; in questo modo, tali campi saranno visualizzati in fondo alla tabella.

La tabella contiene le seguenti colonne:

- **Versione da usare.** Questa colonna visualizza un elenco a discesa che controlla se un campo verrà utilizzato a valle e se usare le trasformazioni suggerite. Per impostazione predefinita, l'elenco rispecchia le raccomandazioni.

Per i predittori ordinari trasformati, l'elenco a discesa contiene tre opzioni: **Trasformata**, **Originale** e **Non utilizzare**.

Per i predittori ordinari non trasformati, le scelte sono: **Originale** e **Non utilizzare**.

Per i campi data/ora derivati e per i predittori creati, le scelte sono: **Trasformata** e **Non utilizzare**.

Per i campi data originali l'elenco a discesa è disattivato e impostato su **Non utilizzare**.

Nota: per i predittori con versioni originale e trasformata, il passaggio dalla versione **Originale** a quella **Trasformata** aggiorna automaticamente le impostazioni **Livello di misurazione** e **Potere predittivo** per tali funzioni.

- **Nome.** Ogni nome di campo è un collegamento. Fare clic su un nome per visualizzare ulteriori informazioni relative al campo nella visualizzazione collegata. Per ulteriori informazioni, consultare l'argomento "Dettagli campo" a pagina 107.
- **Livello di misurazione.** In questa colonna è visualizzata l'icona che rappresenta il tipo di dati; passare il puntatore del mouse sopra l'icona per visualizzare un'etichetta (continuo, ordinale, nominale e così via) che descrive i dati.
- **Potere predittivo.** Il potere predittivo è visualizzato solo per i campi consigliati da ADP. Questa colonna non è visualizzata se non è stato definito un obiettivo. Il potere predittivo è compreso tra 0 e 1, e i valori più elevati indicano predittori "migliori". In generale, il potere predittivo è utile per confrontare i predittori all'interno di un'analisi ADP, ma non deve essere effettuato alcun confronto tra i valori del potere predittivo di analisi diverse.

Riepilogo delle azioni

Per ogni azione svolta dalla preparazione automatica dati, i predittori di input vengono trasformati e/o eliminati tramite filtri; i campi che sopravvivono a un'azione vengono utilizzati nella successiva. I campi che superano tutti i passaggi sono quelli di cui si consiglia l'utilizzo nella modellazione, mentre gli input a predittori trasformati e creati vengono esclusi.

Il Riepilogo delle azioni è una semplice tabella che elenca le azioni di elaborazione svolte da ADP. Se vi è un'**Azione** sottolineata, farvi clic sopra per visualizzare ulteriori dettagli sulle azioni intraprese in una visualizzazione collegata. Per ulteriori informazioni, consultare l'argomento "Dettagli azione" a pagina 108.

Nota: sono visualizzate solo le versioni originale e trasformata definitiva di ciascun campo, non le versioni intermedie utilizzate durante l'analisi.

Potere predittivo

Visualizzato per impostazione predefinita la prima volta che viene eseguita l'analisi o quando si seleziona **Predittori consigliati per l'uso nell'analisi** nella visualizzazione principale Riepilogo elaborazione campi, il grafico mostra il potere predittivo dei predittori consigliati. I campi sono ordinati in base al potere predittivo, a partire dal campo con il valore più elevato.

Per le versioni trasformate dei predittori ordinari, il nome del campo riflette il suffisso scelto nel pannello Nomi di campi della scheda Impostazioni; ad esempio: *_transformed*.

Dopo i singoli nomi dei campi sono visualizzate le icone del livello di misurazione.

Il potere predittivo di ciascun predittore consigliato viene calcolato da una regressione lineare o modello naïve Bayes, in base al fatto che l'obiettivo sia continuo o relativo alla categoria.

Tabella Campi

Visualizzata quando si fa clic su **Obiettivo**, **Predittori** o **Predittori non utilizzati** nella visualizzazione principale Riepilogo elaborazione campi, la visualizzazione Tabella campi mostra una semplice tabella con un elenco delle funzioni pertinenti.

La tabella contiene due colonne:

- **Nome.** Il nome del predittore.

Per gli obiettivi viene utilizzato il nome o l'etichetta originale del campo, anche se l'obiettivo è stato trasformato.

Per le versioni trasformate dei predittori ordinari, il nome riflette il suffisso scelto nel pannello Nomi di campi della scheda Impostazioni; ad esempio: *_transformed*.

Per i campi derivati da date ed ore, viene utilizzato il nome della versione trasformata finale; ad esempio: *bdate_years*.

Per i predittori creati, viene utilizzato il nome del predittore creato; ad esempio: *Predictor1*.

- **Livello di misurazione.** Visualizza l'icona che rappresenta il tipo di dati.

Per l'Obiettivo, il **Livello di misurazione** rispecchia sempre la versione trasformata se l'obiettivo è stato trasformato, ad esempio passando da ordinale (insieme ordinato) a continuo (intervallo, scala) o viceversa.

Dettagli campo

Visualizzata quando si fa clic su un **Nome** nella visualizzazione principale Campi, la visualizzazione Dettagli campo contiene la distribuzione, i valori mancanti e gli eventuali grafici del potere predittivo per il campo selezionato. Inoltre, vengono visualizzati anche la cronologia di elaborazione per il campo ed il nome del campo trasformato (se applicabile).

Per ogni grafico impostato sono visualizzate due versioni affiancate per confrontare il campo con e senza l'applicazione delle trasformazioni; se non esiste una versione trasformata del campo, il grafico viene visualizzato solo per la versione originale. Per i campi data o ora derivati e i predittori creati, i grafici sono visualizzati solo per il nuovo predittore.

Nota: se un campo viene escluso perché dispone di troppe categorie, viene visualizzata solo la cronologia di elaborazione.

Grafico della distribuzione

La distribuzione dei campi continui è rappresentata sotto forma di istogramma a cui è sovrapposta una curva normale e con una linea verticale di riferimento per il valore medio; i campi categoriali sono visualizzati sotto forma di grafico a barre.

Gli istogrammi sono dotati di etichette che mostrano la deviazione standard e l'asimmetria; tuttavia, l'asimmetria non è visualizzata se il numero massimo dei valori è 2 o se la varianza del campo originale è inferiore a 10-20.

Passare il puntatore del mouse sul grafico per visualizzare la media degli istogrammi o il numero e la percentuale sul totale dei record per le per le categorie dei grafici a barre.

Grafico dei valori mancanti

I grafici a torta confrontano la percentuale dei valori mancanti con e senza l'applicazione delle trasformazioni; le etichette del grafico mostrano la percentuale.

Se ADP ha utilizzato la gestione dei valori mancanti, il grafico a torta dopo la trasformazione comprende anche il valore di sostituzione (cioè il valore utilizzato al posto di quelli mancanti) sotto forma di etichetta.

Passare il puntatore del mouse sopra il grafico per visualizzare il numero dei valori mancanti e la percentuale del numero totale di record.

Grafico del potere predittivo

Per i campi consigliati, i grafici a barre mostrano il potere predittivo prima e dopo la trasformazione. Se l'obiettivo è stato trasformato, il potere predittivo calcolato è relativo all'obiettivo trasformato.

Nota: i grafici del potere predittivo non vengono visualizzati se non è definito alcun obiettivo oppure se si fa clic sull'obiettivo nel pannello della visualizzazione principale.

Passare il puntatore del mouse sul grafico per visualizzare il valore del potere predittivo.

Tabella Cronologia elaborazione

La tabella mostra come è stata derivata la versione trasformata di un campo. Le azioni intraprese da ADP sono elencate nell'ordine in cui sono state eseguite; tuttavia, per alcuni passaggi è possibile che siano state intraprese più azioni per un determinato campo.

Nota: questa tabella non viene visualizzata per i campi che non sono stati trasformati.

Le informazioni della tabella sono suddivise in due o tre colonne:

- **Azione.** Il nome dell'azione. Ad esempio, Predittori continui. Per ulteriori informazioni, consultare l'argomento "Dettagli azione".
- **Dettagli.** L'elenco delle procedure eseguite. Ad esempio, Trasforma in unità standard.
- **Funzione.** Viene visualizzata solo per i predittori creati e visualizza la combinazione lineare dei campi di input, ad esempio $.06*age + 1.21*height$.

Dettagli azione

Visualizzata quando si seleziona un'**Azione** sottolineata nella visualizzazione principale Riepilogo delle azioni, la visualizzazione collegata Dettagli azione mostra informazioni generali e specifiche di un'azione per ogni fase di elaborazione effettuata; i dettagli relativi a un'azione specifica sono visualizzati per primi.

Per ogni azione, la descrizione viene utilizzata come titolo nella parte superiore della visualizzazione collegata. I dettagli specifici delle singole azioni sono visualizzati sotto al titolo e possono comprendere il numero di predittori derivati, i campi riformulati, le trasformazioni dell'obiettivo, le categorie unite o riordinate e i predittori creati o esclusi.

A ogni azione, il numero di predittori utilizzato nell'elaborazione può variare, ad esempio a causa dell'esclusione o dell'unione di predittori.

Nota: se un'azione è stata disattivata o se non è stato specificato alcun obiettivo, quando si fa clic sull'azione nella visualizzazione principale Riepilogo delle azioni, viene visualizzato un messaggio di errore invece dei dettagli dell'azione.

Le possibili azioni sono nove, ma non tutte sono necessariamente attive per ogni analisi.

Tabella Campi testo

La tabella mostra il numero di:

- Valori vuoti finali tagliati.
- Predittori esclusi dall'analisi.

Tabella Predittori data e ora

La tabella mostra il numero di:

- Durate derivate da predittori di data e ora.
- Elementi di data e ora.
- Predittori di data e ora derivati in totale.

La data o l'ora di riferimento è visualizzata come nota a piè di pagina se sono state calcolate delle durate delle date.

Tabella Screening dei predittori

La tabella mostra il numero dei seguenti predittori esclusi dall'elaborazione:

- Costanti.
- Predittori con troppi valori mancanti.
- Predittori con troppi casi in un'unica categoria.
- Campi nominali (insiemi) con troppe categorie.
- Predittori esclusi in totale.

Tabella Controlla livello di misurazione

La tabella mostra il numero dei campi riformulati, suddivisi in:

- Campi ordinali (insiemi ordinati) riformulati come campi continui.
- Campi continui riformulati come campi ordinali.
- Numero totale riformulato.

Se nessun campo di input (obiettivo o predittore) è continuo o ordinale, il totale viene visualizzato come un piè di pagina.

Tabella Valori anomali

La tabella mostra il conteggio delle modalità con cui sono stati gestiti i valori anomali.

- Il numero di campi continui per cui i valori anomali sono stati rilevati e tagliati oppure il numero di campi continui per cui i valori anomali sono stati rilevati ed impostati su mancanti, in base alle impostazioni nel pannello Prepara input e obiettivo nella scheda Impostazioni.
- Il numero dei campi continui esclusi perché costanti dopo la gestione dei valori anomali.

Un piè di pagina indica il punto di interruzione dei valori anomali, mentre viene mostrato un altro piè di pagina se nessun campo di input (obiettivo o predittore) è continuo.

Tabella Valori mancanti

La tabella mostra il numero dei campi i cui valori mancanti sono stati sostituiti, suddivisi in:

- Obiettivo. Se non viene specificato alcun obiettivo, questa riga non è visualizzata.
- Predittori. A sua volta suddiviso nel numero di nominali (insieme), ordinali (insieme ordinato) e continui.
- Il numero totale di valori mancanti sostituiti.

Tabella Obiettivo

La tabella indica se l'obiettivo è stato trasformato, illustrato come:

- Trasformazione di Box-Cox alla normalità. Questo valore è ulteriormente suddiviso in colonne che mostrano i criteri specificati (media e deviazione standard) e il valore Lambda.
- Categorie obiettivo riordinate per migliorare la stabilità.

Tabella Predittori categoriali

La tabella mostra il numero di predittori categoriali:

- Le cui categorie sono state riordinate dalla più bassa alla più alta per migliorare la stabilità.
- Le cui categorie sono state unite per aumentare al massimo l'associazione all'obiettivo.
- Le cui categorie sono state unite per gestire le categorie sparse.

- Esclusi a causa di una scarsa associazione all'obiettivo.
- Esclusi perché erano costanti dopo l'unione.

Se non erano presenti predittori categoriali viene visualizzata una nota a piè di pagina.

Tabella Predittori continui

In questo caso le tabelle sono due. La prima visualizza uno dei seguenti numeri di trasformazioni:

- Valori dei predittori trasformati in unità standard. Sono visualizzati inoltre il numero dei predittori trasformati, la media specificata e la deviazione standard.
- Valori dei predittori associati a un intervallo comune. Sono visualizzati inoltre il numero di predittori trasformati mediante una trasformazione Min/Max e i valori minimi e massimi specificati.
- Valori di predittori categorizzati e il numero di predittori categorizzati.

La seconda tabella riporta i dettagli di creazione dello spazio dei predittori, visualizzati sotto forma di numero di predittori:

- Creati.
- Esclusi a causa di una scarsa associazione all'obiettivo.
- Esclusi perché erano costanti dopo la categorizzazione.
- Esclusi perché erano costanti dopo la creazione.

Se l'input non includeva predittori continui viene visualizzata una nota a piè di pagina.

Generazione di un nodo Ricava

Quando si genera un nodo Ricava, esso applica la trasformazione inversa dell'obiettivo al campo punteggio. Per impostazione predefinita, il nodo immette il nome del campo punteggio che sarebbe generato da una funzione di modeler automatico (quale Classificatore automatico o Numerico automatico) o dal nodo dell'insieme. Se è stato trasformato un obiettivo scala (intervallo), il campo punteggio è visualizzato nelle unità trasformate, ad esempio log(\$) anziché \$. Per interpretare e utilizzare questi risultati è necessario riconvertire il valore previsto nella scala originale.

Nota: è possibile generare un nodo Ricava solo quando il nodo ADP contiene un'analisi che ridimensiona un obiettivo intervallo (è selezionato il ridimensionamento Box-Cox nel pannello Prepara input e obiettivo). Non è possibile generare un nodo Ricava se l'obiettivo non è un intervallo o se non è selezionata la modifica della scala di Box-Cox.

Il nodo Ricava viene creato in modalità multipla e utilizza @FIELD nell'espressione; è quindi possibile aggiungere l'obiettivo trasformato, se necessario. Si considerino ad esempio i seguenti dati:

- Nome campo obiettivo: response
- Nome campo obiettivo trasformato: response_transformed
- Nome campo punteggio: \$XR-response_transformed

Il nodo Ricava crea un nuovo campo: \$XR-response_transformed_inverse.

Nota: se non viene utilizzata una funzione di modeler automatico o il nodo dell'insieme, è necessario modificare il nodo Ricava per trasformare il campo del punteggio corretto per il proprio modello.

Target continui normalizzati

Per impostazione predefinita, se viene selezionata la casella di controllo **Ridimensiona target continuo con trasformazione di Box-Cox** nel pannello Prepara input e obiettivo, l'obiettivo viene trasformato e viene creato un nuovo campo che sarà l'obiettivo della creazione del modello. Ad esempio, se l'obiettivo

originale era *response*, il nuovo obiettivo sarà *response_transformed*; i modelli a valle del nodo ADP acquisiranno automaticamente questo nuovo obiettivo.

A seconda dell'obiettivo originale, però, questo può causare dei problemi. Ad esempio, se l'obiettivo era *Età*, i valori del nuovo obiettivo non saranno *Anni*, ma una versione trasformata di *Anni*. Ciò significa che non è possibile esaminare ed interpretare i punteggi, perché non sono espressi in unità riconoscibili. In questo caso si può applicare una trasformazione inversa che riporterà le unità trasformate al formato che avrebbero dovuto avere. A questo scopo, eseguire le operazioni seguenti:

1. Dopo avere fatto clic su **Analizza dati** per eseguire l'analisi ADP, selezionare *Nodo Ricava* dal menu *Genera*.
2. Collocare il nodo Ricava dopo il nugget nell'area del modello.

Il nodo Ricava ripristinerà le dimensioni originali del campo punteggio in modo tale che la previsione sia espressa nei valori *Anni* originali.

Per impostazione predefinita, il nodo Ricava trasforma il campo punteggio generato da una funzione di modeler automatico o da un modello di classificazione binario. Se si crea un modello singolo è necessario modificare il nodo Ricava per derivare il campo dal campo punteggio effettivo. Se si desidera valutare il modello, occorre aggiungere l'obiettivo trasformato al campo **Deriva da** del nodo Ricava. In questo modo si applica la stessa trasformazione inversa all'obiettivo e tutti i nodi Valutazione o Analisi a valle useranno i dati trasformati correttamente a condizione che i nodi vengano modificati per utilizzare nomi di campi anziché metadati.

Se si desidera anche ripristinare il nome originale, è possibile utilizzare un nodo Filtro per rimuovere il campo obiettivo originale, se ancora presente, e ridenominare i campi obiettivo e punteggio.

Nodo Tipo

È necessario specificare le proprietà dei campi in un nodo origine o in un nodo Tipo distinto. Questa funzione è analoga in entrambi i nodi. Sono disponibili le seguenti proprietà:

- **Campo.** Fare doppio clic sul nome di un campo per specificare valore ed etichette di campo per i dati in IBM SPSS Modeler. Per esempio, qui è possibile visualizzare o modificare i metadati dei campi importati da IBM SPSS Statistics. Analogamente, è possibile creare nuove etichette per campi e loro valori. Le etichette specificate qui vengono visualizzate in IBM SPSS Modeler in base alle selezioni effettuate nella finestra di dialogo Proprietà stream.
- **Misurazione.** Livello di misurazione utilizzato per descrivere le caratteristiche dei dati in un campo specifico. Se tutti i dettagli di un campo sono noti, il campo viene definito **completamente istanziato**. Per ulteriori informazioni, consultare l'argomento "Livelli di misurazione" a pagina 112.
Nota: il livello di misurazione di un campo è diverso dal relativo tipo di archiviazione, che indica se i dati sono archiviati come stringhe, interi, numeri reali, date, ore o timestamp.
- **Valori.** Questa colonna consente di specificare le opzioni di lettura dei valori dei dati dall'insieme di dati. L'opzione **Specifici** permette invece di specificare i livelli di misurazione e i valori in una finestra di dialogo a parte. È anche possibile scegliere di passare i campi senza leggerne i valori. Per ulteriori informazioni, consultare l'argomento "Valori dei dati" a pagina 115.
- **Mancante/i.** Utilizzata per specificare come vengono gestiti i valori mancanti del campo. Per ulteriori informazioni, consultare l'argomento "Definizione dei valori mancanti" a pagina 118.
- **Controllo.** In questa colonna è possibile impostare opzioni per assicurarsi che i valori dei campi siano inclusi nei valori o nell'intervallo specificato. Per ulteriori informazioni, consultare l'argomento "Controllo dei valori del tipo" a pagina 119.
- **Ruolo.** Proprietà utilizzata per indicare ai nodi Modelli se i campi verranno trattati come campi di **Input** (campi predittori) o **Obiettivo** (campi previsti) nel processo di apprendimento automatico. **Entrambi** e **Nessuno** sono altri ruoli possibili, oltre a **Partizione**, che indica un campo utilizzato per eseguire la partizione dei record in campioni separati per addestramento, test e convalida. Il valore

Suddivisione specifica che i modelli separati verranno costruiti per ciascun valore possibile del campo. Per ulteriori informazioni, consultare l'argomento "Impostazione del ruolo del campo" a pagina 119.

La finestra del nodo Tipo permette di specificare ulteriori opzioni:

- Quando un nodo Tipo è stato istanziato (tramite specifiche manuali, lettura di valori o esecuzione del flusso) è possibile utilizzare il pulsante del menu di strumenti per scegliere **Ignora campi unici**. Questa opzione consente di ignorare i campi contenenti un solo valore.
- Mediante il pulsante del menu di strumenti è inoltre possibile scegliere **Ignora insieme grandi** per un nodo Tipo istanziato. In questo modo verranno automaticamente ignorati insieme con un numero elevato di membri.
- Mediante il pulsante del menu Strumenti è possibile scegliere **Converti interi continui in ordinali** per un nodo Tipo istanziato. Per ulteriori informazioni, consultare l'argomento "Conversione di dati continui" a pagina 114.
- Mediante il pulsante del menu di strumenti è infine possibile generare un nodo Filtro per scartare i campi selezionati.
- Utilizzando i pulsanti con l'icona degli occhiali, è possibile specificare Leggi o Passa come impostazione di default per tutti i campi. La scheda Tipi nel nodo origine passa i campi per default, mentre il nodo Tipo legge i valori per default.
- Tramite il pulsante **Cancella valori** è possibile annullare le modifiche apportate in questo nodo ai valori dei campi (valori non ereditati) e leggere di nuovo i valori mediante operazioni upstream. Questa opzione è utile per reimpostare le modifiche apportate a monte di campi specifici.
- Tramite il pulsante **Cancella tutti i valori** è possibile reimpostare i valori per **tutti** i campi letti nel nodo. Questa opzione consente di impostare la colonna *Valori* su **Leggi** per tutti i campi, pertanto risulta utile per reimpostare i valori di tutti i campi e leggere nuovamente valori e tipi mediante operazioni upstream.
- Mediante il menu di scelta rapida è possibile scegliere **Copia** per copiare gli attributi di un campo per un altro campo. Per ulteriori informazioni, consultare l'argomento "Copia degli attributi del tipo" a pagina 120.
- Mediante l'opzione **Visualizza impostazioni campi inutilizzati** è possibile visualizzare le impostazioni del tipo relative ai campi non più inclusi nei dati o non più connessi al nodo Tipo corrente. Questa opzione è utile quando si riutilizza un nodo Tipo per insiemi di dati che sono stati modificati.

Livelli di misurazione








Il livello di misurazione (già "tipo di dati" o "tipo di utilizzo") descrive l'utilizzo dei campi di dati in IBM SPSS Modeler. Il livello di misurazione si può specificare nella scheda Tipi di un nodo origine o Tipo. Per esempio, è possibile impostare il livello di misurazione per un campo intero con valori 1 e 0 su *Flag*. In genere, questo indica che 1=*Vero* e 0=*Falso*.

Classe di archiviazione e misurazione. Si noti che il livello di misurazione di un campo è diverso dal tipo di archiviazione del campo, in quanto l'archiviazione indica se i dati vengono archiviati come stringhe, numeri interi, reali, date, ore o timestamp. Mentre i tipi di dati possono essere modificati in qualsiasi punto di un flusso mediante un nodo Tipo, l'archiviazione deve essere stabilita in fase di input quando si leggono i dati in IBM SPSS Modeler (anche se in seguito è possibile modificarla tramite una funzione di conversione). Per ulteriori informazioni, consultare l'argomento "Impostazione dell'archiviazione e della formattazione dei campi" a pagina 24.

Alcuni nodi Modelli indicano, tramite icone sulla loro scheda Campi, i tipi di livello di misurazione consentito per i rispettivi campi di input e obiettivo.

Icone livello di misurazione

Tabella 15. Icone livello di misurazione

Icona	Livello di misurazione
	Default
	Continuo
	Categoriale
	Indicatore
	Nominale
	Ordinale
	Typeless

Sono disponibili i seguenti livelli di misurazione:

- **Default.** I dati di cui non si conosce il tipo di archiviazione né i valori (per esempio perché non sono ancora stati letti) sono visualizzati come <Default>.
- **Continuo.** Utilizzato per descrivere i valori numerici, come un intervallo 0–100 o 0.75–1.25. Un valore continuo può essere un numero intero, un numero reale o un valore di data/ora.
- **Categoriali.** Utilizzato per valori stringa nel caso in cui non sia noto un numero esatto di valori distinti. Questo è un tipo di dati **non istanziato**, ovvero, tutte le informazioni possibili sull'archiviazione e l'utilizzo dei dati non sono ancora note. Dopo la lettura dei dati, il livello di misurazione sarà definito come *Indicatore*, *Nominale* o *Senza tipo*, a seconda del numero massimo di membri per i campi nominali specificato nella finestra di dialogo Proprietà flusso.
- **Indicatore.** Utilizzato per i dati con due valori distinti che indicano la presenza o l'assenza di una caratteristica, come true e false, Sì e No oppure 0 e 1. I valori utilizzati possono variare, ma uno deve essere sempre indicato come valore "true" e l'altro come valore "false". I dati possono essere rappresentati come testo, numero intero, numero reale, data, ora o timestamp.
- **Nominale.** Utilizzato per descrivere i dati con più valori distinti, ognuno dei quali viene trattato come membro di un insieme, per esempio small/medium/large. I dati nominali possono avere qualsiasi classe di archiviazione —numerica, stringa o data/ora. Tenere presente che l'impostazione del livello di misurazione su *Nominale* non cambia automaticamente i valori sul tipo di archiviazione stringa.
- **Ordinale.** Utilizzato per descrivere i dati con più valori distinti che hanno un ordine intrinseco. Per esempio, le categorie di stipendio o le classificazioni del livello di soddisfazione possono essere immesse come dati ordinali. L'ordine è definito dal criterio di ordinamento naturale degli elementi di dati. Per esempio, 1, 3, 5 è il criterio di ordinamento di default per un insieme di interi, mentre HIGH, LOW, NORMAL (ordine alfabetico crescente) è l'ordine per un insieme di stringhe. Il livello di misurazione ordinale consente di definire un insieme di dati categoriali come dati ordinali per la visualizzazione, la creazione del modello e l'esportazione ad altre applicazioni (per esempio IBM SPSS Statistics) che sono in grado di riconoscere i dati ordinali come un tipo distinto. È possibile utilizzare un campo ordinale in tutti i casi in cui è possibile utilizzare un campo nominale. Inoltre, è possibile definire come ordinali i campi con qualsiasi tipo di classe di archiviazione (reale, intero, stringa, data, ora e così via).
- **Senza tipo.** Utilizzato per dati che non rientrano in nessuno dei tipi sopra riportati, per i campi con un solo valore o per dati nominali in cui l'insieme ha un numero di membri superiore al massimo definito. È utile anche nei casi in cui il livello di misurazione potrebbe corrispondere a un insieme con un numero eccessivo di membri (per esempio un numero di account). Quando si seleziona **Senza tipo** per un campo, il ruolo viene impostato automaticamente su **Nessuno**, con **ID Record** come unica

alternativa. La dimensione massima di default per gli insiemi è 250 valori univoci. Questo numero può essere adeguato o disattivato nella scheda Opzioni di nella finestra di dialogo Proprietà flusso, a cui si può accedere dal menu Strumenti.

È possibile specificare i livelli di misurazione manualmente, oppure consentendo al software di leggere i dati e di stabilirne il livello di misurazione in base ai valori letti.

In alternativa, quando si dispone di diversi campi di dati continui che devono essere trattati come dati categoriali, è possibile scegliere un'opzione per convertirli. Per ulteriori informazioni, consultare l'argomento "Conversione di dati continui".

Per utilizzare l'impostazione automatica

1. In un nodo Tipo o nella scheda Tipi di un nodo origine, impostare la colonna *Valori* su **<Leggi>** nei campi desiderati. I metadati saranno così disponibili per tutti i nodi downstream. È possibile impostare rapidamente tutti i campi su **<Leggi>** o **<Passa>** utilizzando i pulsanti con l'icona degli occhiali nella finestra di dialogo.
2. Per leggere subito i valori dalla sorgente dati, fare clic su **Leggi valori**.

Per impostare manualmente il livello di misurazione di un campo

1. Selezionare un campo nella tabella.
2. Dall'elenco a discesa nella colonna *Misurazione*, selezionare un livello di misurazione per il campo.
3. In alternativa, utilizzare Ctrl-A o Ctrl-clic per selezionare più campi prima di selezionare un livello di misurazione dall'elenco a discesa.

Conversione di dati continui

Se i dati relativi alla categoria vengono considerati continui, ciò potrebbe influire sulla qualità di un modello, specialmente nel caso di un campo obiettivo; ad esempio, la creazione di un modello di regressione invece di un modello binario. Per evitare che questo accada è possibile convertire intervalli di numeri interi in tipi categoriali come *Ordinale* o *Flag*.

1. Dal pulsante del Menu di strumenti per operazioni e generazioni (con il simbolo dello strumento), selezionare **Converti interi continui in ordinali**. Viene visualizzata la finestra di dialogo dei valori di conversione.
2. Specificare le dimensioni dell'intervallo che verrà automaticamente convertito; questa operazione si applica a tutti gli intervalli inferiori e uguali alle dimensioni immesse.
3. Fare clic su **OK**. Gli intervalli interessati vengono convertiti in *Flag* o *Ordinale* e visualizzati nella scheda Tipi del nodo Tipo.

Risultati della conversione

- Quando un campo *Continuo* con tipo di archiviazione numero intero viene modificato in *Ordinale*, i valori inferiore e superiore vengono espansi in modo da includere tutti i valori interi dal valore più basso a quello più alto. Per esempio, se l'intervallo è 1, 5, l'insieme di valori è 1, 2, 3, 4, 5.
- Se il campo *Continuo* viene trasformato in *Flag*, i valori inferiore e superiore diventano i valori falso e vero del campo Flag.

Che cos'è la creazione istanza?

Per **creazione istanza** si intende il processo di lettura o definizione di informazioni quali il tipo di archiviazione e i valori per un campo dati. Per ottimizzare le risorse del sistema, l'istanziamento è un processo diretto dall'utente — l'utente indica al software di leggere i valori specificando le opzioni nella scheda Tipi in un nodo di origine oppure eseguendo i dati mediante un nodo Tipo.

- I dati con tipi sconosciuti vengono anche detti **non istanziati**. I dati il cui tipo di archiviazione e i cui valori sono sconosciuti vengono visualizzati nella colonna *Misurazione* della scheda Tipi come **<Default>**.

- Quando si dispone delle informazioni relative al tipo di archiviazione di un campo, stringa o numerico, i dati vengono definiti **parzialmente istanziati**. I livelli **Categoriale** o **Continuo** sono livelli di misurazione parzialmente istanziati. **Categoriale** specifica, per esempio, che il campo è simbolico, ma non se sia nominale, ordinale o flag.
- Quando sono noti tutti i dettagli relativi ad un tipo, inclusi i valori, viene visualizzato un livello di misurazione **completamente istanziato** — nominale, ordinale, flag o continuo — in questa colonna.
Nota: il tipo *continuo* viene utilizzato per i campi di dati parzialmente istanziati e completamente istanziati. I dati continui possono essere numeri interi o numeri reali.

Durante l'esecuzione di un flusso di dati con un nodo Tipo, i tipi non istanziati vengono parzialmente istanziati in base ai valori dei dati iniziali. Dopo essere passati per il nodo, tutti i dati vengono completamente istanziati, a meno che i valori non siano stati impostati su **<Passa>**. Se l'esecuzione viene interrotta, i dati rimarranno parzialmente istanziati. Una volta istanziata la scheda Tipi, i valori di un campo sono statici in quel punto del flusso. Ciò significa che le modifiche a monte non influenzeranno i valori di un particolare campo, nemmeno se si riesegue il flusso. Per cambiare o aggiornare i valori in base ai nuovi dati o a ulteriori manipolazioni, è necessario modificarli nella scheda Tipi o impostare il valore del campo desiderato su **<Leggi>** o **<Leggi +>**.

Informazioni sull'istanziamento

In genere, se l'insieme di dati non è di grandi dimensioni e non si prevede di aggiungere campi nel flusso in seguito, l'istanziamento a livello del nodo origine risulta il metodo più vantaggioso. L'istanziamento in un nodo Tipo distinto risulta tuttavia utile nei seguenti casi:

- L'insieme di dati è di grandi dimensioni e il flusso filtra un sottoinsieme precedente al nodo Tipo.
- I dati sono stati filtrati nel flusso.
- I dati sono stati uniti o accodati nel flusso.
- Durante l'elaborazione vengono derivati nuovi campi di dati.

Valori dei dati

Tramite la colonna *Valori* della scheda Tipi è possibile leggere automaticamente i valori dai dati o specificare livelli di misurazione e valori in una finestra di dialogo separata.

Le opzioni disponibili dall'elenco a discesa Valori forniscono le istruzioni per l'impostazione automatica, come illustrato nella seguente tabella.

Tabella 16. Istruzioni per l'impostazione automatica

Opzione	Funzione
<Leggi>	Selezionare questa opzione per leggere i valori durante l'esecuzione del nodo.
<Leggi+>	I dati vengono letti e accodati ai dati correnti (se presenti).
<Passa>	Non verrà letto alcun dato.
<Corrente>	Vengono mantenuti i valori dei dati correnti.
Specifica...	Viene aperta una finestra di dialogo separata in cui specificare le opzioni su valori e livelli di misurazione.

Se si esegue un nodo Tipo o si fa clic su **Leggi valori**, i valori vengono impostati automaticamente e letti dalla sorgente dati in base alla selezione. I valori possono essere specificati anche manualmente tramite l'opzione Specifica o facendo doppio clic su una cella nella colonna *Campo*.

Dopo aver apportato le modifiche per i campi nel nodo Tipo, è possibile utilizzare i seguenti pulsanti nella barra degli strumenti della finestra di dialogo per reimpostare le informazioni sui valori:

- Tramite il pulsante **Cancella valori** è possibile annullare le modifiche apportate in questo nodo ai valori dei campi (valori non ereditati) e leggere di nuovo i valori mediante operazioni upstream. Questa opzione è utile per reimpostare le modifiche apportate a monte di campi specifici.
- Tramite il pulsante **Cancella tutti i valori** è possibile reimpostare i valori per **tutti** i campi letti nel nodo. Questa opzione consente di impostare la colonna *Valori* su **Leggi** per tutti i campi, pertanto risulta utile per reimpostare i valori di tutti i campi e leggere nuovamente valori e livelli di misurazione mediante operazioni upstream.

Utilizzo della finestra di dialogo Valori

Se si fa clic sulla colonna *Valori* o *Mancante* nella scheda Tipi, viene visualizzato un elenco a discesa di valori predefiniti. Se si sceglie l'opzione *Specifica* nell'elenco, verrà aperta una finestra di dialogo distinta nella quale è possibile impostare le opzioni per la lettura, la definizione, l'assegnazione delle etichette e la gestione dei valori del campo selezionato.

Molti controlli sono comuni a tutti i tipi di dati. Tali controlli comuni vengono illustrati di seguito.

Misurazione. Visualizza il livello di misurazione selezionato. È possibile modificare l'impostazione per indicare il modo in cui si intende utilizzare i dati. Se un campo chiamato *giorno_della_settimana* contiene, per esempio, numeri che rappresentano singoli giorni, è possibile che si desideri modificare questa impostazione in dati nominali per creare un nodo di distribuzione che esamini ciascuna categoria individualmente.

Archiviazione. Visualizza il tipo di archiviazione, se noto. Il livello di misurazione prescelto non influisce sui tipi di archiviazione. Per modificare il tipo di archiviazione è possibile utilizzare la scheda Dati nei nodi origine Testo fisso e Testo variabile o una funzione di conversione in un nodo Riempimento.

Campo del modello. Per i campi generati in seguito al calcolo del punteggio di un nugget del modello è possibile visualizzare anche i dettagli del campo del modello. Essi comprendono il nome del campo obiettivo, nonché il ruolo del campo nella modellazione (valore previsto, probabilità, propensione, ecc.).

Valori. Selezionare un metodo di lettura dei valori per il campo selezionato. Le selezioni eseguite in questa finestra sovrascrivono quelle eseguite in precedenza dalla colonna *Valori* della finestra di dialogo del nodo Tipo. Le scelte per la lettura dei valori includono:

- **Leggi dai dati.** Selezionare questa opzione per leggere i valori durante l'esecuzione del nodo. L'opzione è analoga a <Leggi>.
- **Passa.** Selezionare questa opzione per non leggere i dati per il campo corrente. L'opzione è analoga a <Passa>.
- **Specifica valori ed etichette.** Le opzioni disponibili consentono di specificare i valori e le etichette per il campo selezionato. Utilizzata insieme al controllo valori, questa opzione permette di specificare i valori in base alle informazioni note sul campo corrente, inoltre attiva controlli univoci per ciascun tipo di campo. Le singole opzioni relative ai valori e alle etichette verranno illustrate negli argomenti successivi. *Nota:* non è possibile specificare valori o etichette per un campo il cui livello di misurazione è *Typeless* o <Default>.
- **Estendi valori dai dati.** Selezionare questa opzione per accodare ai dati correnti i valori sopra specificati. Se, per esempio, il campo *campo_1* include un intervallo (0,10) e si immette un intervallo di valori (8,16), l'intervallo viene esteso aggiungendo 16, senza rimuovere il minimo originale. Il nuovo intervallo sarà (0,16). La scelta di questa opzione imposta automaticamente l'opzione di impostazione automatica su <Leggi+>.

Controlla valori. Selezionare un metodo per forzare i valori in modo che siano conformi ai valori continui, flag o nominali specificati. Questa opzione corrisponde alla colonna *Controllo* nella finestra di dialogo del nodo Tipo, le cui impostazioni vengono sovrascritte da quelle specificate in questa finestra. Se utilizzato insieme all'opzione *Specifica valori*, il controllo dei valori consente di uniformare i valori nei dati ai valori previsti. Per esempio, se si specificano valori come 1, 0 e si utilizza l'opzione **Scarta**, è possibile scartare tutti i record con valori diversi da 1 o 0.

Definisci vuoti. Selezionare questa opzione per attivare i controlli sotto riportati, che consentono di dichiarare valori mancanti o vuoti nei dati.

- **Tabella Valori mancanti.** Consente di definire valori specifici (per esempio 99 o 0) come vuoti. Il valore deve essere appropriato per il tipo di archiviazione del campo.
- **Intervallo.** Utilizzato per specificare un intervallo di valori mancanti, ad esempio le età 1–17 o maggiori di 65. Se un valore limite viene lasciato vuoto, l'intervallo non avrà limite; ad esempio, se viene specificato il limite inferiore 100 senza limite superiore, tutti i valori maggiori o uguali a 100 verranno definiti come mancanti. I valori limite sono inclusi nell'intervallo. Per esempio, un intervallo con limite inferiore 5 e limite superiore 10 includerà i valori 5 e 10. È possibile definire un intervallo di valori mancanti per qualsiasi tipo di classe di archiviazione, inclusi data/ora e stringa. In quest'ultimo caso, per stabilire se un valore rientra nell'intervallo verrà utilizzato il criterio di ordinamento alfabetico.
- **Null/Spazio vuoto.** È inoltre possibile definire come vuoti i valori di sistema **null** (visualizzati nei dati come \$null\$) e **spazio vuoto** (valori stringa senza caratteri visibili). Ai fini dell'analisi, anche il nodo Tipo considera le stringhe vuote come spazi vuoti, nonostante vengano archiviate in maniera diversa internamente e possano, in determinati casi, essere gestite diversamente.

Nota: per codificare i vuoti come non definiti o \$null\$, è necessario utilizzare il nodo Riempimento.

Descrizione. Utilizzare questa casella di testo per specificare un'etichetta del campo. Le etichette vengono visualizzate in numerose posizioni, per esempio nei grafici, nelle tabelle, nell'output e nei browser di modelli, in base alle selezioni eseguite nella finestra di dialogo delle proprietà del flusso.

Definizione di valori ed etichette per dati continui

Il livello di misurazione *Continuo* si utilizza per i campi numerici. Per i dati continui sono disponibili tre tipi di archiviazione:

- Reale
- Intero
- Data/ora

Per modificare tutti i campi continui si utilizza la stessa finestra di dialogo; il tipo di archiviazione è visualizzato solo a scopo informativo.

Definizione dei valori

I controlli seguenti sono specifici dei campi continui e consentono di definire un intervallo di valori:

Ultimo. Specificare un limite inferiore per l'intervallo dei valori.

Primo. Specificare un limite superiore per l'intervallo di valori.

Definizione delle etichette

È possibile definire le etichette per qualsiasi valore di un campo intervallo. Fare clic sul pulsante **Etichette** per aprire una finestra di dialogo nella quale è possibile definire le etichette dei valori.

Finestra secondaria dei valori e delle etichette

Se si fa clic su **Etichette** nella finestra di dialogo Valori relativa a un campo intervallo, verrà aperta una nuova finestra di dialogo nella quale è possibile definire le etichette relative a qualsiasi valore compreso nell'intervallo.

Per definire le coppie di valori ed etichette, è possibile utilizzare le colonne *Valori* e *Etichette* della tabella. Sono visualizzate le coppie già definite. Per aggiungere nuove coppie, è possibile fare clic su una cella

vuota e immettere un valore e la relativa etichetta. *Nota:* l'aggiunta di coppie valore/valore-etichetta alla tabella non comporterà l'aggiunta di nuovi valori al campo, ma verranno semplicemente creati i metadati per il valore del campo.

Le etichette specificate nel nodo Tipo vengono visualizzate in numerose posizioni (come testi descrittivi, etichette dell'output e così via), in base alle selezioni eseguite nella finestra di dialogo delle proprietà del flusso.

Definizione di valori ed etichette per dati nominali e ordinali

I livelli di misurazione nominali (insieme) e ordinali (insieme ordinato) indicano che i valori dei dati vengono utilizzati in modo discreto come un membro dell'insieme. I tipi di archiviazione per un insieme possono essere: stringa, numero intero, numero reale o data/ora.

I controlli seguenti sono specifici dei campi nominali e ordinali e consentono di definire i valori e le etichette:

Valori. La colonna *Valori* della tabella consente di specificare i valori in base alle informazioni note sul campo corrente. Tramite questa tabella è possibile immettere i valori previsti per il campo e quindi controllare la conformità dell'insieme di dati a questi valori utilizzando l'elenco a discesa Controlla valori. I pulsanti freccia e quello contrassegnato da una X consentono di modificare l'ordine dei valori esistenti o di eliminarli.

Etichette. La colonna *Etichette* consente di specificare le etichette per ogni valore del campo insieme. Le etichette vengono visualizzate in numerose posizioni, per esempio nei grafici, nelle tabelle, nell'output e nei browser di modelli, in base alle selezioni eseguite nella finestra di dialogo delle proprietà del flusso.

Definizione dei valori relativi a un campo flag

I campi flag vengono utilizzati per visualizzare i dati con due valori distinti. I tipi di archiviazione per i flag possono essere: stringa, numero intero, numero reale o data/ora.

Vero. Specificare un valore flag per il campo quando la condizione è soddisfatta.

Falso. Specifica un valore flag per il campo quando la condizione non viene soddisfatta.

Etichette. Specifica le etichette per ogni valore del campo flag. Le etichette vengono visualizzate in numerose posizioni, per esempio nei grafici, nelle tabelle, nell'output e nei browser di modelli, in base alle selezioni eseguite nella finestra di dialogo delle proprietà del flusso.

Definizione dei valori mancanti

La colonna **Mancante** della scheda Tipi indica se per un campo è stata definita la gestione dei valori mancanti. Le impostazioni possibili sono:

On (*). Indica che per questo campo è definita la gestione dei valori mancanti. Questo è anche possibile utilizzando il nodo Riempimento a valle oppure tramite una specifica esplicita utilizzando l'opzione Specifica (vedere qui di seguito).

Off. Il campo non ha una gestione dei valori mancanti definita.

Specifica. Scegliere questa opzione per visualizzare una finestra di dialogo in cui è possibile dichiarare che i valori espliciti vengano considerati come valori mancanti per questo campo.

Controllo dei valori del tipo

Se si attiva l'opzione *Controllo* per ogni campo, verranno esaminati tutti i valori nel singolo campo in modo da stabilire se rispondono alle impostazioni correnti del tipo o ai valori specificati nella finestra di dialogo *Specifica valori*. Questa opzione è utile per ripulire gli insiemi di dati e ridurne le dimensioni con un'unica operazione.

L'impostazione della colonna *Controllo* nella finestra di dialogo del nodo *Tipo*, definisce le operazioni che verranno eseguite nel caso venga individuato un valore che non rientra nei limiti del tipo. Per cambiare le impostazioni *Controllo* per un campo, utilizzare l'elenco a discesa relativo a tale campo nella colonna *Controllo*. Per definire le impostazioni *Controllo* per tutti i campi, fare clic nella colonna *Campo* e premere Ctrl-A. Quindi utilizzare l'elenco a discesa per ogni campo nella colonna *Controllo*.

Sono disponibili le seguenti impostazioni *Controllo*:

Nessuna. I valori verranno passati senza effettuare alcun controllo. (impostazione predefinita).

Annulla. Cambia i valori esterni ai limiti nel valore null di sistema (\$null\$).

Forza. Verranno controllati i campi i cui livelli di misurazione sono completamente istanziati, per scoprire se contengono valori che non rientrano negli intervalli specificati. I valori non specificati verranno convertiti in un valore valido per quel livello di misurazione, applicando le seguenti regole:

- Per i flag, tutti i valori diversi da quelli veri e falsi vengono convertiti nel valore falso.
- Per gli insiemi (nominali o ordinali), qualsiasi valore sconosciuto viene convertito nel primo membro dei valori dell'insieme.
- I numeri maggiori del limite superiore di un intervallo vengono sostituiti con il limite superiore.
- I numeri inferiori al limite inferiore di un intervallo vengono sostituiti con il limite inferiore.
- Ai valori null in un intervallo viene assegnato il valore intermedio dell'intervallo.

Scarta. Quando vengono trovati valori non validi, viene scartato l'intero record.

Avvisa. Viene conteggiato il numero di elementi non validi e il risultato del conteggio viene riportato nella finestra di dialogo delle proprietà del flusso al termine della lettura di tutti i dati.

Termina. Il primo valore non valido trovato determina l'interruzione dell'esecuzione del flusso. L'errore viene riportato nella finestra di dialogo delle proprietà del flusso.

Impostazione del ruolo del campo

Il ruolo di un campo specifica il modo in cui il campo viene utilizzato nella creazione del modello — ad esempio, se si tratta di un campo di input o obiettivo (oggetto della previsione).

Nota: i ruoli *Partizione*, *Frequenza* e *ID record* possono essere applicati solo ad un campo singolo.

Sono disponibili i ruoli seguenti:

Input. Il campo verrà utilizzato come input per l'apprendimento automatico (campo predittore).

Obiettivo. Il campo verrà utilizzato come output o come obiettivo per l'apprendimento automatico (uno dei campi che il modello tenterà di prevedere).

Entrambi. Il campo verrà utilizzato sia come input che come output dal nodo *Apriori*. Tutti gli altri nodi *Modelli* ignoreranno il campo.

Nessuna. Il campo verrà ignorato dall'apprendimento automatico. I campi il cui livello di misurazione è impostato su **Senza tipo** vengono automaticamente impostati su **Nessuno** nella colonna *Ruolo*.

Partizione. Indica un campo utilizzato per partizionare i dati in campioni di addestramento, di test e di convalida (facoltativo) distinti. Questo campo deve essere un tipo di insieme istanziato con due o tre valori possibili (definiti nella finestra di dialogo Valori di campo). Il primo valore rappresenta il campione di addestramento, il secondo rappresenta il campione di test e il terzo (se presente) rappresenta il campione di convalida. Tutti i valori aggiuntivi vengono ignorati e non è possibile utilizzare i campi flag. Si noti che, per utilizzare la partizione in un'analisi, è necessario attivare il partizionamento nella scheda Opzioni modello del nodo di creazione del modello o di analisi appropriato. Se il partizionamento è attivato, i record con valori null del campo partizione vengono esclusi dall'analisi. Se nel flusso sono stati definiti più campi partizione, nella scheda Campi di ogni nodo Modelli applicabile è necessario selezionare un singolo campo partizione. Se nei dati non esiste già un campo appropriato, è possibile crearne uno utilizzando un nodo Partizione o Ricava. Per ulteriori informazioni, consultare l'argomento "Nodo Partizione" a pagina 146.

Suddivisione. (Solo per campi nominali, ordinali e flag) Specifica che deve essere creato un modello per ciascun valore possibile del campo.

Frequenza. (Solo per campi numerici) L'impostazione di questo ruolo consente di utilizzare un valore del campo come fattore di peso della frequenza per il record. Questa funzione è supportata solo dai nodi C&R Tree, CHAID, QUEST e Modelli lineari; tutti gli altri nodi ignorano questo ruolo. Il peso della frequenza si attiva mediante l'opzione **Utilizza campo frequenza** nella scheda Campi dei nodi Modelli che supportano la funzione.

ID record. Il campo viene utilizzato come identificatore univoco del record. Questa funzione è ignorata dalla maggior parte dei nodi, ma è supportata dai modelli lineari ed è richiesta per i nodi IBM Netezza per il mining nel database.

Copia degli attributi del tipo

È facile copiare gli attributi del tipo di un campo (per esempio valori, opzioni di controllo e valori mancanti) e incollarli in un altro campo:

1. Fare clic con il pulsante destro del mouse sul campo di cui si desidera copiare gli attributi.
2. Dal menu di scelta rapida scegliere **Copia**.
3. Fare clic con il pulsante destro del mouse sul campo o sui campi di cui si desidera cambiare gli attributi.
4. Dal menu di scelta rapida, selezionare **Incolla speciale**. *Nota:* è possibile selezionare più campi utilizzando il metodo Ctrl-clic o l'opzione **Seleziona campi** dal menu di scelta rapida.

Verrà aperta una nuova finestra di dialogo dalla quale selezionare gli attributi specifici che si desidera incollare. Se gli attributi vengono incollati in più campi, le opzioni selezionate in questa finestra verranno applicate a tutti i campi obiettivo.

Incolla i seguenti attributi. Selezionare dall'elenco sotto riportato l'attributo che si desidera incollare in un altro campo.

- **Tipo.** Selezionare per incollare il livello di misurazione.
- **Valori.** Selezionare per incollare i valori del campo.
- **Mancante.** Selezionare per incollare le impostazioni del valore mancante.
- **Controllo.** Selezionare per incollare le opzioni di verifica del valore.
- **Ruolo.** Selezionare per incollare il ruolo di un campo.

Scheda per l'impostazione dei formati dei campi

Nella scheda Formato dei nodi Tabella e Tipo sono visualizzati l'elenco dei campi correnti o inutilizzati e le opzioni di formattazione relativi a ogni campo. Di seguito è riportata la descrizione di ogni colonna disponibile nella tabella di formattazione dei campi:

Campo. Mostra il nome del campo selezionato.

Formato. Se si fa doppio clic su una cella della colonna, verrà aperta una finestra di dialogo nella quale è possibile specificare la formattazione dei singoli campi. Per ulteriori informazioni, consultare l'argomento "Impostazione delle opzioni relative al formato dei campi". La formattazione specificata in questo punto sostituisce la formattazione specificata nelle proprietà generali del flusso.

Nota: i nodi Esporta Statistics ed Output Statistics esportano file *.sav* che includono la formattazione per i campi nei propri metadati. Se è specificato un formato per i campi non supportato dal formato di file *.sav* di IBM SPSS Statistics, il nodo utilizzerà il formato predefinito di IBM SPSS Statistics.

Giustifica. Utilizzare questa colonna per specificare le modalità di giustificazione dei valori nella colonna della tabella. L'impostazione di default è **Automatico** e giustifica a sinistra i valori simbolici e a destra i valori numerici. Per ignorare l'impostazione di default, selezionare **A sinistra**, **A destra** o **Al centro**.

Larghezza colonna. Per default, le larghezze delle colonne vengono calcolate automaticamente in base ai valori dei campi. Per ignorare il calcolo automatico della larghezza, fare clic su una cella della tabella e selezionare una larghezza dall'elenco a discesa. Per immettere una larghezza personalizzata non presente nell'elenco, visualizzare la finestra di dialogo secondaria relativa ai formati dei campi facendo doppio clic su una cella della tabella nella colonna Campo o Formato. In alternativa, fare clic con il pulsante destro del mouse su una cella e selezionare **Imposta formato**.

Visualizza campi correnti. Per default, la finestra di dialogo mostra l'elenco dei campi attivi. Per visualizzare l'elenco dei campi inutilizzati, selezionare **Visualizza impostazioni campi inutilizzati**.

Menu di scelta rapida. Il menu di scelta rapida per questa scheda offre numerose opzioni di selezione e aggiornamento delle impostazioni. Fare clic con il tasto destro del mouse in una colonna per visualizzare tale menu.

- **Seleziona tutto.** Seleziona tutti i campi.
- **Seleziona nessuno.** Cancella la selezione.
- **Seleziona campi.** Seleziona i campi in base al tipo o alle caratteristiche di archiviazione. Le opzioni sono **Seleziona categoriali**, **Seleziona continui** (numerici), **Seleziona senza tipo**, **Seleziona stringhe**, **Seleziona numeri** o **Seleziona data/ora**. Per ulteriori informazioni, consultare l'argomento "Livelli di misurazione" a pagina 112.
- **Imposta formato.** Apre una finestra secondaria in cui è possibile specificare la data, l'ora e le opzioni decimali per ognuno dei campi.
- **Imposta giustificato.** Imposta la giustificazione per il campo o i campi selezionati. Le opzioni sono **Automatico**, **Al centro**, **A sinistra** o **A destra**.
- **Imposta larghezza colonna.** Imposta la larghezza dei campi selezionati. Specificare **Automatica** per leggere il valore dai dati oppure impostare il valore su 5, 10, 20, 30, 50, 100 o 200.

Impostazione delle opzioni relative al formato dei campi

La formattazione dei campi viene definita in una finestra di dialogo secondaria della scheda Formato dei nodi Tipo e Tabella. Se prima di aprire questa finestra sono stati selezionati più campi, verranno utilizzate le impostazioni del primo campo selezionato. Se si fa clic su **OK** dopo avere impostato le opzioni, tali opzioni verranno applicate a tutti i campi selezionati nella scheda Formato.

Per ognuno dei campi sono disponibili le opzioni seguenti, molte delle quali possono anche essere impostate nella finestra di dialogo delle proprietà del flusso. Le impostazioni eseguite a livello di campo hanno tuttavia la precedenza sulle impostazioni di default definite per il flusso.

Formato data. Selezionare un formato di data da utilizzare per i campi di archiviazione delle date o quando le stringhe vengono interpretate come date dalle funzioni di data di CLEM.

Formato ora. Selezionare un formato di ora da utilizzare per i campi di archiviazione dell'ora o per le stringhe interpretate come orari dalle funzioni di ora CLEM.

Formato di visualizzazione numeri. È possibile scegliere il formato di visualizzazione standard (####.###), scientifico (#.###E+##) o valuta (\$###.##).

Simbolo decimale. Selezionare una virgola (,) o un punto (.) come separatore decimale.

Simbolo di raggruppamento. Per i formati di visualizzazione dei numeri, selezionare il simbolo utilizzato per raggruppare i valori (per esempio la virgola per 3.000,00). Le opzioni disponibili sono Nessuno, Punto, Virgola, Spazio e Definito in base alle opzioni internazionali (in questo caso verrà utilizzato il valore di default per le opzioni internazionali correnti).

Numero di decimali (Standard, Scientifico, Valuta, Esporta). Per i formati di visualizzazione dei numeri, specifica il numero di cifre decimali da utilizzare quando si visualizzano i numeri reali. Questa opzione viene specificata separatamente per ogni formato di visualizzazione.

Giustifica. Specifica le modalità di giustificazione dei valori nella colonna. L'impostazione di default è **Automatico** e giustifica a sinistra i valori simbolici e a destra i valori numerici. Per ignorare l'impostazione di default, selezionare A sinistra, A destra o Al centro.

Larghezza colonna. Per default, le larghezze delle colonne vengono calcolate automaticamente in base ai valori dei campi. È possibile specificare una larghezza personalizzata in base a intervalli di cinque, utilizzando le frecce a destra della casella di riepilogo.

Applicazione di filtri o ridenominazione di campi

È possibile rinominare o escludere i campi in qualsiasi punto di un flusso. Per esempio, è possibile che un ricercatore medico non sia interessato al livello di potassio (dati a livello di campo) dei pazienti (dati a livello di record). Grazie a questo nodo il ricercatore può filtrare il campo *K* (potassio). Si può eseguire questa operazione mediante un nodo Filtro separato o tramite la scheda Filtro di un nodo origine o di output. La funzionalità è identica, indipendentemente dal nodo da cui vi si accede.

- Da nodi origine quali Testo variabile, Testo fisso, File Statistics e XML, è possibile rinominare o filtrare i campi mentre i dati vengono letti in IBM SPSS Modeler.
- Un nodo Filtro consente di rinominare o filtrare i campi in qualsiasi punto del flusso.
- Dai nodi Esporta Statistics, Trasformazioni Statistics, Modelli Statistics e Output Statistics è possibile filtrare o rinominare i campi per adeguarli agli standard di denominazione di IBM SPSS Statistics. Per ulteriori informazioni, consultare l'argomento "Ridenominazione o filtro dei campi per IBM SPSS Statistics" a pagina 350.
- In tutti i nodi sopracitati è possibile utilizzare la scheda Filtro per definire o modificare gli insiemi di risposte multiple. Per ulteriori informazioni, consultare l'argomento "Modifica di insiemi di risposte multiple" a pagina 124.
- Infine, il nodo Filtro può essere utilizzato per mappare i campi tra nodi origine.

Impostazione delle opzioni di filtro

Nella tabella utilizzata nella scheda Filtro, per ogni campo viene riportato il nome del campo in entrata nel nodo e in uscita dal nodo. È possibile utilizzare le opzioni in questa tabella per rinominare o filtrare i campi duplicati o non necessari per operazioni a valle.

- **Campo.** Visualizza i campi di input dalle origini dati attualmente connesse.
- **Filtro.** Visualizza lo stato del filtro di tutti i campi di input. I campi filtrati sono contrassegnati da una X rossa in questa colonna e non verranno passati a valle. Fare clic nella colonna *Filtro* per attivare e disattivare il filtro di un campo selezionato. È anche possibile selezionare le opzioni per più campi contemporaneamente, tramite il metodo di selezione Maiusc-clic.

- **Campo.** Visualizza i campi all'uscita dal nodo Filtro. I nomi duplicati sono visualizzati in rosso. È possibile modificare i nomi dei campi facendo clic su questa colonna e immettendo il nuovo nome. Per rimuovere i campi, fare clic nella colonna *Filtro* per disattivare i campi duplicati.

Tutte le colonne nella tabella possono essere ordinate facendo clic sulla relativa intestazione.

Visualizza campi correnti. Selezionare per visualizzare i campi per gli insiemi di dati connessi in modo attivo al nodo Filtro. Questa opzione è selezionata per default e rappresenta il metodo di utilizzo dei nodi Filtro più comune.

Visualizza impostazioni campi inutilizzati. Selezionare per visualizzare i campi per gli insiemi di dati non più connessi al nodo Filtro. Questa opzione risulta utile quando si copiano nodi Filtro da un flusso in un altro stream e quando si ricaricano nodi Filtro.

Menu del pulsante Filtro

Fare clic sul pulsante Filtro nell'angolo superiore sinistro della finestra di dialogo per accedere a un menu contenente numerose scelte rapide e altre opzioni.

È possibile scegliere di:

- Rimuovere tutti i campi.
- Includere tutti i campi.
- Attivare/disattivare tutti i campi.
- Rimuovere i duplicati. *Nota:* se si seleziona questa opzione, verranno rimosse tutte le occorrenze del nome duplicato, inclusa la prima.
- Rinominare i campi e gli insiemi di risposte multiple per conformarli a quelli delle altre applicazioni. Per ulteriori informazioni, consultare l'argomento "Ridenominazione o filtro dei campi per IBM SPSS Statistics" a pagina 350.
- Troncare i nomi dei campi.
- Anonimizzare i nomi dei campi e degli insiemi di risposte multiple.
- Utilizzare i nomi dei campi di input.
- Modificare gli insiemi di risposte multiple. Per ulteriori informazioni, consultare l'argomento "Modifica di insiemi di risposte multiple" a pagina 124.
- Impostare lo stato di default dei filtri.

È anche possibile utilizzare i pulsanti con le frecce nella parte superiore della finestra di dialogo per includere o scartare i campi per default. Questa opzione risulta utile per insiemi di dati di grandi dimensioni in cui per le operazioni a valle è necessario includere solo alcuni campi. Per esempio, è possibile selezionare solo i campi da conservare e specificare che tutti gli altri devono essere scartati (anziché selezionare i campi da scartare uno per uno).

Troncamento dei nomi di campo

Dal menu del pulsante Filtro (nell'angolo superiore sinistro della scheda Filtro) è possibile scegliere di troncare i nomi di campo.

Lunghezza massima. Consente di specificare un numero di caratteri per limitare la lunghezza dei nomi di campo.

Numero di cifre. Se i nomi di campo non sono più univoci dopo il troncamento, verranno ulteriormente troncati e differenziati mediante l'aggiunta di cifre. È possibile specificare il numero di cifre utilizzate. I pulsanti con le frecce permettono di modificare il numero.

Ad esempio, la tabella riportata di seguito mostra il modo in cui i nomi dei campi in un insieme di dati medici vengono troncati utilizzando le impostazioni di default (lunghezza massima=8 e numero di cifre=2).

Tabella 17. Troncamento del nome del campo

Nomi di campi	Nomi di campo troncati
InputPaziente1	InputP01
InputPaziente2	InputP02
Battito Cardiaco	BattitoC
Pressione	Pressione

Anonimizzazione dei nomi di campi

È possibile anonimizzare i nomi dei campi da qualunque nodo contenente una scheda Filtro facendo clic sul menu del pulsante Filtro nell'angolo superiore sinistro e selezionando **Anonimizza nomi di campi**. I nomi dei campi anonimizzati consistono in un prefisso sotto forma di stringa e in un valore su base numerica univoco.

Anonimizza nomi di. Scegliere **Solo campi selezionati** per anonimizzare solo i nomi dei campi già selezionati nella scheda Filtro. L'impostazione di default è **Tutti i campi**, che anonimizza i nomi di tutti i campi.

Prefisso dei nomi dei campi. Il prefisso di default per i nomi di campi anonimizzati è **anon_**; se si preferisce utilizzare un prefisso diverso, scegliere **Personalizzato** e digitare il prefisso.

Anonimizza insiemi di risposte multiple. Anonimizza i nomi degli insiemi di risposte multiple analogamente a quanto avviene per i campi. Per ulteriori informazioni, consultare l'argomento "Modifica di insiemi di risposte multiple".

Per ripristinare i nomi di campi originali, scegliere **Utilizza nomi campi di input** dal menu di pulsanti di filtro.

Modifica di insiemi di risposte multiple

È possibile aggiungere o modificare gli insiemi di risposte multiple da qualunque nodo contenente una scheda Filtro facendo clic sul menu del pulsante Filtro nell'angolo superiore sinistro e selezionando **Modifica insiemi di risposte multiple**.

Gli insiemi di risposte multiple vengono utilizzati per registrare i dati che possono avere più di un valore per ciascun caso — ad esempio, quando agli intervistati viene chiesto quali musei sono stati visitati oppure le riviste lette. Gli insiemi di risposte multiple possono essere importati in IBM SPSS Modeler mediante un nodo origine di Data Collection o un nodo origine Statistics e possono essere definiti in IBM SPSS Modeler mediante un nodo Filtro.

Fare clic su **Nuovo** per creare un nuovo insieme di risposte multiple, oppure su **Modifica** per modificare un insieme esistente.

Nome ed Etichetta. Riportano il nome e la descrizione dell'insieme.

Tipo. Le domande a risposta multipla possono essere gestite in due diversi modi:

- **Insieme dicotomie multiple.** Viene creato un campo flag separato per ogni risposta possibile: in questo modo, se le riviste sono 10, vi saranno 10 campi flag, ciascuno dei quali può avere dei valori quali 0 o 1 per indicare *vero* o *falso*. Il valore conteggiato consente di specificare quale valore viene conteggiato come vero. Questo metodo è utile quando si desidera permettere agli intervistati di scegliere tutte le opzioni pertinenti.

- **Insieme di categorie multiple.** Viene creato un campo nominale per ogni risposta fino al numero massimo di risposte date da un determinato intervistato. Ogni campo nominale contiene valori che rappresentano le risposte possibili, quali 1 per *Time*, 2 per *Newsweek* e 3 per *PC Week*. Questo metodo è molto utile quando si desidera limitare il numero di risposte — ad esempio, quando agli intervistati viene richiesto di scegliere le tre riviste lette con maggior frequenza.

Campi nell'insieme. Utilizzare le icone a destra per aggiungere o rimuovere i campi.

Commenti

- Tutti i campi compresi in un insieme di risposte multiple devono avere lo stesso tipo di archiviazione.
- Gli insiemi e i campi in essi contenuti sono due entità separate. Ad esempio, eliminando un insieme, i campi inclusi non vengono eliminati — vengono eliminati solo i collegamenti tra tali campi. L'insieme risulterà ancora visibile a monte del punto di eliminazione, ma non a valle.
- Se i campi vengono ridenominati utilizzando un nodo Filtro (direttamente nella scheda oppure selezionando le opzioni Rinomina per IBM SPSS Statistics, **Tronca** o **Anonimizza** nel menu Filtro), verranno aggiornati anche tutti i riferimenti a tali campi utilizzati negli insiemi di risposte multiple. I campi di un insieme di risposte multiple eliminati mediante il nodo Filtro non saranno invece rimossi dall'insieme di risposte multiple. Tali campi, sebbene non più visibili nel flusso, continuano ad avere dei riferimenti nell'insieme di risposte multiple; questo potrebbe essere un aspetto da tenere presente quando si esegue un'esportazione, per esempio.

Nodo dell'insieme

Il nodo dell'insieme combina due o più nugget del modello per ottenere previsioni più precise di quelle ricavabili dai singoli modelli. Se si combinano le previsioni di più modelli è possibile superare le limitazioni dei singoli modelli e ottenere un livello di precisione complessiva più elevato. I modelli combinati in questo modo forniscono in genere prestazioni almeno equivalenti a quelle del singolo modello migliore, quando non addirittura superiori.

Questa combinazione di nodi viene eseguita automaticamente nei nodi Modelli automatici Classificatore automatico, Numerico automatico e Cluster automatico.

Dopo avere utilizzato il nodo dell'insieme è possibile utilizzare un nodo Analisi o Valutazione per confrontare la precisione dei risultati combinati con quelli di ciascun modello di input. Per eseguire questa operazione, verificare che l'opzione **Filtra campi generati dai modelli di classificazione binari** sia deselezionata nella scheda Impostazioni del nodo dell'insieme.

Campi di output

Ogni nodo dell'insieme genera un campo contenente i punteggi combinati. Il nome del campo è composto dal campo obiettivo specificato con il prefisso $\$XF_$, $\$XS_$ o $\$XR_$, a seconda del livello di misurazione del campo - rispettivamente, flag, nominale (insieme) o continuo (intervallo). Per esempio, se il campo obiettivo è un campo flag denominato *risposta*, il campo di output sarà $\$XF_{risposta}$.

Campi di confidenza o propensione. Per i campi nominali e flag, vengono creati ulteriori campi di confidenza o propensione in base al metodo dell'insieme, come indicato in dettaglio nella tabella riportata di seguito.

Tabella 18. Creazione campo del metodo dell'insieme.

Metodo dell'insieme	Nome del campo
Voting Confronto ponderato con confidenza Confronto ponderato con propensione grezza Confronto ponderato con propensione regolata Confidenza più elevata vince	$\$XFC_{<campo>}$

Tabella 18. Creazione campo del metodo dell'insieme (Continua).

Metodo dell'insieme	Nome del campo
Propensione grezza media	\$XFRP_<campo>
Propensione grezza/regolata media	\$XFAP_<campo>

Impostazioni del nodo dell'insieme

Campo obiettivo per insieme. Selezionare un campo singolo che sarà utilizzato come obiettivo da due o più modelli a monte. I modelli a monte possono utilizzare target flag, nominali o continui, ma per poter combinare i punteggi è necessario che almeno due modelli condividano lo stesso obiettivo.

Filtra campi generati dai modelli di classificazione binari. Elimina dall'output tutti i campi aggiuntivi generati dai singoli modelli che alimentano il nodo dell'insieme. Selezionare questa casella di controllo se si è interessati unicamente al punteggio combinato di tutti i modelli di input. Verificare che questa opzione sia deselezionata se, per esempio, si desidera utilizzare un nodo Analisi o Valutazione per confrontare la precisione del punteggio combinato con quella dei singoli modelli di input.

Le impostazioni disponibili dipendono dal livello di misurazione del campo selezionato come obiettivo.

Target continui

Per i target continui viene calcolata la media dei punteggi. Questo è l'unico metodo disponibile per combinare i punteggi.

Quando si esegue la media di punteggi o stime, il nodo dell'insieme utilizza un calcolo degli errori standard per stabilire la differenza tra i valori misurati o stimati e quelli veri, e per mostrare la corrispondenza di queste stime. I calcoli degli errori standard vengono generati di default per i nuovi modelli; tuttavia, è possibile deselezionare la casella di controllo per i modelli esistenti, per esempio se devono essere rigenerati.

Obiettivi categoriali

Per gli obiettivi categoriali sono supportati numerosi metodi fra cui il **confronto**, che consiste nel conteggio del numero di volte in cui ogni possibile valore previsto viene scelto e nella selezione del valore con il totale più elevato. Per esempio, se tre modelli su cinque prevedono *sì* e gli altri due *no*, il *sì* vince per un voto di 3 a 2. In alternativa, i voti possono essere **ponderati** in base al valore di confidenza o di propensione di ogni previsione. Vengono quindi sommati i pesi e viene nuovamente selezionato il valore con il totale più alto. La confidenza della previsione finale è la somma dei pesi del valore vincente divisa per il numero dei modelli compresi nell'insieme.

Tutti i campi categoriali. Sono supportati i seguenti metodi sia per i campi flag che per i campi nominali:

- Voting
- Confronto ponderato con confidenza
- Confidenza più elevata vince

Solo i campi flag. Solo per i campi flag, sono disponibili anche numerosi metodi basati sulla propensione:

- Confronto ponderato con propensione grezza
- Confronto ponderato con propensione regolata
- Propensione grezza media
- Propensione regolata media

Confronti a pari merito. Per i metodi del confronto è possibile specificare le modalità di risoluzione delle situazioni di pari merito.

- **Selezione casuale.** Viene scelto a caso uno dei valori pari merito.
- **Confidenza più elevata.** Vince il valore pari merito previsto con la confidenza più elevata. Si noti che questa non corrisponde necessariamente alla confidenza più elevata di tutti i valori previsti.
- **Propensione grezza o regolata (solo campi flag).** Il valore pari merito previsto con la massima propensione assoluta, dove la propensione assoluta è calcolata come:

$$\frac{\text{abs}(0.5 - \text{propensione})}{2}$$

O, nel caso della propensione regolata:

$$\text{abs}(0.5 - \text{propensione regolata}) * 2$$

Nodo Ricava

Una delle funzioni più efficaci di IBM SPSS Modeler è rappresentata dalla possibilità di modificare i valori dei dati e derivare nuovi campi dai dati esistenti. Nel corso di lunghi progetti di data mining è normale eseguire diverse derivazioni, per esempio estrarre gli ID clienti da una stringa di dati di registro Web oppure creare un valore del profitto futuro prodotto dal cliente in base ai dati demografici e alle transazioni. Queste trasformazioni possono essere eseguite tramite una grande varietà di nodi Operazioni su campi.

Alcuni nodi consentono di derivare nuovi campi:



Il nodo Ricava modifica valori di dati o crea nuovi campi da uno o più campi esistenti. Crea campi di tipo Formula, Flag, Nominale, Stato, Conteggio e Condizionale.



Il nodo Ricodifica trasforma un insieme di valori categoriali in un altro. L'operazione di ricodifica consente di comprimere categorie o raggruppare dati per l'analisi.



Il nodo Discretizza crea automaticamente nuovi campi nominali (insieme) basati sui valori di uno o più campi continui (intervallo numerico) esistenti. Per esempio, è possibile trasformare un campo continuo relativo al reddito in campo categoriale contenente gruppi di reddito come deviazioni dalla media. Dopo aver creato bin per il nuovo campo, è possibile generare un nodo Ricava basato sui punti di divisione.



Il nodo Crea flag crea campi flag in base ai valori categoriali di uno o più campi nominali.



Il nodo Riorganizza converte un campo nominale o flag in un gruppo di campi in cui è possibile inserire i valori di un altro campo. Per esempio, dato un campo denominato *tipo di pagamento*, con valori di *credito*, *contanti* e *debito*, verrebbero creati tre nuovi campi (*credito*, *contanti*, *debito*), ognuno dei quali può contenere il valore del pagamento effettuato.



Il nodo Cronologia crea nuovi campi contenenti dati dei campi di record precedenti. I nodi Cronologia sono utilizzati in genere per dati sequenziali, per esempio per dati di serie temporali. Prima di utilizzare un nodo Cronologia, può essere utile ordinare i dati con un nodo Ordina.

Utilizzo del nodo Ricava

Il nodo Ricava consente di creare sei tipi di nuovi campi da uno o più campi esistenti:

- **Formula.** Il nuovo campo è il risultato di un'espressione CLEM arbitraria.
- **Flag.** Il nuovo campo è un flag, che rappresenta una condizione specificata.
- **Nominale.** Il nuovo campo è nominale, ovvero i suoi membri sono un gruppo di valori specificati.
- **Stato.** Il nuovo campo è uno di due stati. La commutazione da uno stato all'altro è attivata da una condizione specificata.
- **Conteggio.** Il nuovo campo si basa sul numero di volte in cui una condizione è risultata vera.
- **Condizionale.** Il nuovo campo è il valore di una di due espressioni, in base al valore di una condizione.

Ogni nodo contiene un insieme di opzioni speciali nella finestra di dialogo del nodo Ricava. Tali opzioni verranno illustrate negli argomenti seguenti.

Impostazione delle opzioni di base relative al nodo Crea variabile

Nella parte superiore della finestra di dialogo per i nodi Crea variabile sono incluse diverse opzioni per la selezione del tipo di nodo Crea variabile desiderato.

Modalità. Selezionare **Singola** o **Multipla** per derivare un solo campo o più campi. Se si seleziona **Multipla**, la finestra di dialogo cambia per includere opzioni per più campi.

Deriva campo. Per nodi Ricava con modalità singola, specificare il nome del campo che si desidera derivare e aggiungere a ciascun record. Il nome di default è RicavaN, dove N è il numero di nodi Ricava creati fino a questo momento nella sessione corrente.

Deriva come. Selezionare un tipo di nodo Ricava, per esempio Formula o Nominale, dall'elenco a discesa. Per ogni tipo viene creato un nuovo campo in base alle condizioni specificate nella finestra di dialogo relativa al tipo.

La selezione di un'opzione dall'elenco a discesa comporta l'aggiunta di un nuovo insieme di controlli alla finestra di dialogo principale, a seconda delle proprietà di ogni tipo di nodo Ricava.

Tipo campo. Selezionare un livello di misurazione, per esempio continuo, categoriale o flag, per il nodo derivato. Questa opzione è comune a tutte le forme di nodi Ricava.

Nota: per derivare nuovi campi è spesso necessario utilizzare funzioni speciali o espressioni matematiche. Il generatore di espressioni, disponibile nella finestra di dialogo per tutti i tipi di nodi Ricava, semplifica la creazione di tali espressioni e rende disponibile il controllo delle regole e un elenco completo delle espressioni CLEM.

Derivazione di campi multipli

L'impostazione della modalità **Multipla** in un nodo Crea variabile permette di derivare più campi basati sulla stessa condizione all'interno dello stesso nodo. Questa funzione consente di risparmiare tempo quando si desidera eseguire trasformazioni identiche per più campi dell'insieme di dati. Se si vuole, per esempio, generare un modello di regressione per la previsione del salario corrente in base al salario iniziale e all'esperienza precedente, potrebbe essere utile applicare una trasformazione logaritmica a tutte e tre le variabili asimmetriche. Invece di aggiungere un nuovo nodo Ricava per ogni trasformazione, è

possibile applicare contemporaneamente la stessa funzione a tutti i campi. È sufficiente selezionare tutti i campi dai quali si desidera derivare un nuovo campo, quindi digitare l'espressione di derivazione utilizzando la funzione @FIELD all'interno delle parentesi del campo.

Nota: la funzione @FIELD è uno strumento importante per la derivazione contemporanea di più campi, in quanto consente di fare riferimento al contenuto dei campi correnti senza specificare il nome esatto del campo. Per esempio, un'espressione CLEM utilizzata per applicare una trasformazione logaritmica a più campi è $\log(@FIELD)$.

Se si seleziona la modalità **Multipla**, alla finestra di dialogo vengono aggiunte le seguenti opzioni:

Deriva da. Utilizzare il Selettore di campo per selezionare i campi dai quali derivare nuovi campi. Per ogni campo selezionato verrà generato un campo di output. *Nota:* i campi selezionati non devono necessariamente essere dello stesso tipo di archiviazione. Tuttavia, se la condizione non è valida per *tutti* i campi l'operazione di derivazione non verrà eseguita.

Estensione nome campo. Digitare l'estensione da aggiungere ai nuovi nomi di campo. Se un nuovo campo contiene, per esempio, il registro di *Salario Corrente*, è possibile aggiungere l'estensione *_registro* al nome del campo, ottenendo *Salario Corrente_registro*. Utilizzare i pulsanti di scelta rapida per aggiungere l'estensione come prefisso (all'inizio) o come suffisso (alla fine) del nome di campo. Il nome di default è *RicavaN*, dove *N* è il numero di nodi Ricava creati fino a questo momento nella sessione corrente.

Come nel nodo Ricava in modalità singola, è necessario generare l'espressione da utilizzare per la derivazione di un nuovo campo. Le opzioni disponibili per creare una condizione variano a seconda del tipo di operazione di derivazione selezionata. Tali opzioni verranno illustrate negli argomenti seguenti. Per generare un'espressione, è possibile digitare l'espressione direttamente nei campi della formula oppure utilizzare il Generatore di espressioni facendo clic sul pulsante del calcolatore. Ricordarsi di utilizzare la funzione @FIELD quando ci si riferisce a manipolazioni su più campi.

Selezione di più campi

Per tutti i nodi che eseguono operazioni su più campi di input, per esempio Ricava (modalità multipla), Aggregazione, Ordina, Multiplot e Grafico tempo è possibile selezionare facilmente più campi tramite la finestra di dialogo Seleziona campi.

Ordina per. Per visualizzare i campi disponibili in un determinato ordine, sono disponibili le seguenti opzioni:

- **Naturale.** Visualizza l'ordine dei campi secondo la modalità di passaggio a valle nel flusso di dati nel nodo corrente.
- **Nome.** Utilizza l'ordine alfabetico per ordinare i campi per la visualizzazione.
- **Tipo.** Visualizza i campi ordinati in base al relativo livello di misurazione. Questa opzione è utile quando si selezionano campi con un determinato livello di misurazione.

Selezionare i campi dall'elenco uno per volta, oppure utilizzare i metodi Maiusc-clic e Ctrl-clic per selezionare più campi contemporaneamente. È anche possibile utilizzare i pulsanti nella parte inferiore dell'elenco per selezionare gruppi di campi in base al livello di misurazione, oppure per selezionare o deselegionare tutti i campi nella tabella.

Impostazione delle opzioni relative al nodo Crea variabile formula

I nodi Crea variabile formula creano un nuovo campo per ciascun record in un insieme di dati, in base ai risultati di un'espressione CLEM. Si noti che questa espressione non può essere condizionale. Per derivare valori in base a un'espressione condizionale, utilizzare il tipo flag o condizionale del nodo Ricava.

Formula. Specificare una formula utilizzando il linguaggio CLEM per derivare un valore per il nuovo campo.

Impostazione delle opzioni relative al nodo Ricava indicatore

I nodi Ricava flag vengono utilizzati per indicare una condizione specifica, per esempio pressione sanguigna elevata o inattività dell'account cliente. Per ciascun record viene creato un campo flag e, se la condizione è vera, nel campo viene aggiunto il valore flag per vero.

Valore vero. Specificare il valore da includere nel campo flag per i record che soddisfano la condizione specificata sotto. L'impostazione di default è V.

Valore falso. Specificare il valore da includere nel campo flag per i record che *non* soddisfano la condizione specificata sotto. L'impostazione di default è F.

Valore vero quando. Specificare una condizione CLEM per valutare determinati valori di ciascun record e assegnare al record un valore vero o falso (definito sopra). Si noti che il valore vero verrà assegnato ai record nel caso di valori numerici non falsi.

Nota: per restituire una stringa vuota, digitare virgolette di apertura e di chiusura senza immettere alcun carattere tra di esse (""). Le stringhe vuote vengono utilizzate spesso, per esempio come valore falso, per dare maggior risalto ai valori veri presenti in una tabella. Le virgolette dovrebbero essere utilizzate anche se si desidera ottenere un valore stringa che altrimenti verrebbe trattato come numero

Esempio

Nelle release di IBM SPSS Modeler precedenti alla 12.0, venivano importate più risposte in un singolo campo, con i valori separati da virgole. Ad esempio:

```
museum_of_design,institute_of_textiles_and_fashion  
museum_of_design  
archeological_museum  
$null$  
national_art_gallery,national_museum_of_science,other
```

Per preparare i dati per l'analisi è possibile utilizzare la funzione `hassubstring` per generare un campo flag separato per ciascuna risposta con un'espressione quale:

```
hassubstring(museums,"museum_of_design")
```

Impostazioni delle opzioni del nodo nuovo campo nominale

I nodi Ricava nominale vengono utilizzati per eseguire una serie di condizioni CLEM per determinare la condizione soddisfatta da ciascun record. Ogni volta che viene rispettata una condizione per ciascun record, al nuovo campo derivato verrà aggiunto un valore indicante l'insieme di condizioni soddisfatte.

Valore di default. Specificare il valore da utilizzare nel nuovo campo se nessuna delle condizioni viene soddisfatta.

Imposta campo su. Specificare il valore da immettere nel nuovo campo quando viene soddisfatta una determinata condizione. A ogni valore dell'elenco è associata una condizione, specificata nella colonna adiacente.

Se questa condizione è vera. Specificare una condizione per ogni membro nel campo insieme da elencare. Utilizzare il Generatore di espressioni per selezionare funzioni e campi disponibili. È possibile utilizzare i pulsanti con le frecce e quello contrassegnato da una X per riordinare o rimuovere le condizioni.

Una condizione esegue la verifica dei valori di un particolare campo nell'insieme di dati. Al termine della verifica di ogni condizione, i valori specificati sopra verranno assegnati al nuovo campo in modo da indicare quale condizione è stata rispettata. Se non viene soddisfatta alcuna condizione, verrà utilizzato il valore di default.

Impostazione delle opzioni relative al nodo Ricava stato

I nodi Ricava stato sono simili ai nodi Ricava indicatore. Un nodo Ricava indicatore imposta i valori in base al modo in cui viene soddisfatta una *singola* condizione del record corrente, invece un nodo Ricava stato è in grado di modificare i valori di un campo in base al modo in cui vengono soddisfatte *due condizioni indipendenti*. Ciò significa che il valore viene modificato (attivato o disattivato) quando viene soddisfatta ciascuna condizione.

Stato iniziale. Selezionare se assegnare inizialmente il valore **Attivato** o **Disattivato** a ciascun record del nuovo campo. Si noti che questo valore può cambiare quando viene soddisfatta ciascuna condizione.

Valore "Attivato". Specificare il valore per il nuovo campo quando viene soddisfatta la condizione Attivato.

Imposta su "Attivato" se. Specificare una condizione CLEM per modificare lo stato in Attivato se la condizione è vera. Fare clic sul pulsante del calcolatore per aprire il Generatore di espressioni.

Valore "Disattivato". Specificare il valore per il nuovo campo quando viene soddisfatta la condizione Disattivato.

Imposta su "Disattivato" se. Specificare una condizione CLEM per modificare lo stato in Disattivato se la condizione è falsa. Fare clic sul pulsante del calcolatore per aprire il Generatore di espressioni.

Nota: per specificare una stringa vuota, digitare virgolette di apertura e di chiusura senza immettere alcun carattere tra di esse (""). Le virgolette dovrebbero essere utilizzate anche se si desidera ottenere un valore stringa che altrimenti verrebbe trattato come numero.

Impostazione delle opzioni relative al nodo Crea variabile conteggio

Il nodo Crea variabile conteggio è utilizzato per applicare una serie di condizioni ai valori di un campo numerico nell'insieme di dati. Quando ogni condizione viene soddisfatta, il valore del campo conteggio derivato viene aumentato in base a un incremento prestabilito. Questo tipo di nodo Ricava è utile per dati di serie temporali.

Valore iniziale. Imposta un valore utilizzato per il nuovo campo durante l'esecuzione. Il valore iniziale deve essere una costante numerica. Utilizzare i pulsanti con le frecce per aumentare o diminuire il valore.

Incrementa quando. Specificare la condizione CLEM che, se soddisfatta, determina la modifica del valore derivato in base al numero specificato in Incrementa di. Fare clic sul pulsante del calcolatore per aprire il Generatore di espressioni.

Incrementa di. Imposta il valore utilizzato per incrementare il conteggio. È possibile utilizzare una costante numerica oppure il risultato di un'espressione CLEM.

Reimposta quando. Specificare una condizione che, se soddisfatta, determinerà la nuova impostazione del valore derivato sul valore iniziale. Fare clic sul pulsante del calcolatore per aprire il Generatore di espressioni.

Impostazione delle opzioni relative al nodo Ricava condizionale

I nodi Ricava condizionale utilizzano una serie di istruzioni Se, Allora, Altrimenti per derivare il valore del nuovo campo.

Se. Specifica una condizione CLEM che verrà valutata per ciascun record durante l'esecuzione. Se la condizione è vera (o non falsa, nel caso di numeri), al nuovo campo viene assegnato il valore specificato sotto dall'espressione Allora. Fare clic sul pulsante del calcolatore per aprire il Generatore di espressioni.

Allora. Specifica un valore o un'espressione CLEM per il nuovo campo quando l'istruzione Se sopra specificata è vera (o non falsa). Fare clic sul pulsante del calcolatore per aprire il Generatore di espressioni.

Altrimenti. Specifica un valore o un'espressione CLEM per il nuovo campo quando l'istruzione Se sopra specificata è falsa. Fare clic sul pulsante del calcolatore per aprire il Generatore di espressioni.

Ricodifica di valori con il nodo Ricava

I nodi Ricava possono essere utilizzati anche per ricodificare i valori, per esempio convertendo un campo stringa con valori categoriali in un campo nominale (insieme) numerico.

1. Per Deriva come, selezionare il tipo di campo (nominale, flag e così via) desiderato.
2. Specificare le condizioni per la ricodifica dei valori. Per esempio, è possibile impostare il valore su 1 se Cura='curaA', 2 se Cura='curaB' e così via.

Nodo Riempimento

I nodi Riempimento sono utilizzati per sostituire i valori di campo e modificare la classe di archiviazione. È possibile scegliere di sostituire i valori in base a una condizione CLEM specificata, per esempio @BLANK(@CAMPO). In alternativa, si può scegliere di sostituire tutti i valori null o vuoti con un valore specifico. I nodi Riempimento sono utilizzati spesso insieme al nodo Tipo per sostituire i valori mancanti. Per esempio, è possibile riempire i vuoti con il valore medio di un campo, specificando un'espressione @GLOBAL_MEAN. L'espressione riempirà tutti i vuoti con il valore medio calcolato da un nodo Calcola globali.

Riempi campi Tramite Selettore di campo (pulsante a destra del campo testo), selezionare i campi dall'insieme di dati i cui valori verranno esaminati e sostituiti. Per default, i valori vengono sostituiti in base alle espressioni Condizione e Sostituisci con specificate sotto. Le opzioni Sostituisci riportate sotto consentono di selezionare un metodo di sostituzione alternativo.

Nota: quando si selezionano più campi da sostituire con un valore definito dall'utente, è importante che i tipi di campo siano simili (tutti numerici o tutti simbolici).

Sostituisci. Selezionare uno dei seguenti metodi per sostituire i valori dei campi selezionati:

- **In base alla condizione.** Questa opzione attiva il campo Condizione e il Generatore di espressioni per consentire all'utente di creare un'espressione utilizzata come condizione per effettuare la sostituzione con il valore specificato.
- **Sempre.** Sostituisce tutti i valori del campo selezionato. Per esempio, questa opzione potrebbe essere utilizzata per convertire la classe di archiviazione del reddito in una stringa tramite la seguente espressione CLEM:(to_string(reddito)).
- **Valori vuoti.** Sostituisce tutti i valori vuoti specificati dall'utente nel campo selezionato. Per selezionare i valori vuoti viene utilizzata la condizione standard @BLANK(@CAMPO). *Nota:* per definire i valori vuoti, utilizzare la scheda Tipi del nodo origine o un nodo Tipo.
- **Valori null.** Questa opzione sostituisce tutti i valori null di sistema nel campo selezionato. Per selezionare i valori null viene utilizzata la condizione standard @NULL(@CAMPO).
- **Valori vuoti e null.** Questa opzione sostituisce sia i valori vuoti che null di sistema nel campo selezionato. È utile quando non si è sicuri se i valori null sono stati o meno definiti come valori mancanti.

Condizione. Questa opzione è disponibile quando viene selezionata l'opzione **In base alla condizione**. Utilizzare questa casella di testo per specificare un'espressione CLEM per la valutazione dei campi selezionati. Fare clic sul pulsante del calcolatore per aprire il Generatore di espressioni.

Sostituisci con. Specificare un'espressione CLEM per assegnare un nuovo valore ai campi selezionati. Per sostituire il valore con un valore null, digitare undef nella casella di testo. Fare clic sul pulsante del calcolatore per aprire il Generatore di espressioni.

Nota: quando i campi selezionati sono stringhe, è necessario sostituirli con un valore stringa. L'utilizzo dello 0 di default o di un altro valore numerico come valore sostitutivo per i campi stringa determinerà un errore.

Conversione della classe di archiviazione mediante il nodo Riempimento

La condizione Sostituisci di un nodo Riempimento consente di convertire facilmente la classe di archiviazione per uno o più campi. Per esempio, utilizzando la funzione di conversione `to_integer`, è possibile convertire *reddito* da stringa in numero intero tramite la seguente espressione CLEM: `to_integer(reddito)`.

È possibile visualizzare le funzioni di conversione disponibili e generare automaticamente un'espressione CLEM utilizzando il Generatore di espressioni. Dall'elenco a discesa delle funzioni, selezionare **Conversione** per visualizzare l'elenco delle funzioni di conversione della classe di archiviazione. Sono disponibili le seguenti opzioni di esecuzione:

- `to_integer(ELEMENTO)`
- `to_real(ITEM)`
- `to_number(ITEM)`
- `to_string(ELEMENTO)`
- `to_datetime(ELEMENTO)`
- `to_timestamp(ELEMENTO)`
- `to_date(ELEMENTO)`
- `to_datetime(ELEMENTO)`

Conversione dei valori di data e ora. Si noti che le funzioni di conversione, e qualsiasi altra funzione che richieda un tipo di input specifico quale un valore di data o di ora, dipendono dai formati correnti specificati nella finestra di dialogo Opzioni stream. Per esempio, se si desidera convertire un campo stringa con valori *Gen 2003*, *Feb 2003* e così via, per l'archiviazione dei dati, selezionare **MES AAAA** come formato di data di default per il flusso.

Le funzioni di conversione sono disponibili anche dal nodo Ricava per la conversione temporanea durante un calcolo di derivazione. È inoltre possibile utilizzare il nodo Ricava per eseguire altre manipolazioni, quali ricodifiche di campi stringa con valori categoriali. Per ulteriori informazioni, consultare l'argomento "Ricodifica di valori con il nodo Ricava" a pagina 132.

Nodo Anonimizza

Il nodo Anonimizza consente di mascherare i nomi dei campi, i valori dei campi o entrambi, quando si utilizzano dati da includere in un modello a valle del nodo. In questo modo, il modello generato può essere distribuito liberamente, per esempio al supporto tecnico, senza rischiare che i dati riservati, quali record dei dipendenti o anamnesi dei pazienti, possano essere visualizzati da utenti non autorizzati.

A seconda del punto del flusso in cui il nodo Anonimizza viene situato, potrebbe essere necessario apportare modifiche ad altri nodi. Per esempio, se si inserisce un nodo Anonimizza a monte di un nodo Seleziona, i criteri di selezione di quest'ultimo dovranno essere modificati se agiscono su valori che sono stati anonimizzati.

Il metodo da utilizzare per l'anonimizzazione dipende da diversi fattori. Per i nomi dei campi e tutti i valori di campo ad eccezione di Livelli di misurazione continui, i dati vengono sostituiti da una stringa con il seguente formato:

prefix_Sn

dove *prefix_* è una stringa specificata dall'utente o la stringa di default *anon_* e *n* è un valore intero che parte da 0 e viene incrementato per ciascun valore univoco (per esempio, *anon_S0*, *anon_S1* e così via).

I valori di campo di tipo Continuo devono essere trasformati, poiché gli intervalli numerici si riferiscono a valori interi o reali e non a stringhe. In tal senso, possono essere anonimizzati solo mediante la trasformazione dell'intervallo in un intervallo diverso, vale a dire, mascherando i dati originali. La trasformazione di un valore *x* nell'intervallo viene eseguita nel modo seguente:

$$A*(x + B)$$

dove:

A è un fattore di scala che deve essere maggiore di 0.

B è un offset di traslazione da aggiungere ai valori.

Esempio

Nel caso di un campo *ETÀ* in cui il fattore di scala *A* è impostato su 7 e l'offset di traslazione *B* è impostato su 3, i valori di *ETÀ* vengono trasformati in:

$$7*(ETÀ + 3)$$

Impostazione delle opzioni relative al nodo Anonimizza

A questo punto è possibile scegliere i campi i cui valori dovranno risultare mascherati a valle.

Si noti che i campi di dati devono essere istanziati a monte del nodo Anonimizza prima che possano essere eseguite le operazioni di anonimizzazione. È possibile istanziare i dati facendo clic sul pulsante **Leggi valori** in un nodo tipologia oppure nella scheda Tipi di un nodo di origine.

Campo. Elenca i campi nel dataset corrente. Se alcuni dei campi sono già stati anonimizzati, i relativi nomi anonimizzati vengono visualizzati qui.

Misurazione. Il livello di misurazione del campo.

Anonimizza valori. Selezionare uno o più campi, fare clic sulla colonna e scegliere **Sì** per anonimizzare il valore dei campi utilizzando il prefisso di default **anon_**; scegliere **Specifica** per visualizzare una finestra di dialogo in cui immettere il proprio prefisso o, nel caso dei valori di campi di tipo *Continuo*, specificare se per la trasformazione dei valori dei campi è necessario utilizzare valori casuali o specificati dall'utente. Si noti che i tipi di campo *Continuo* e non *Continuo* non possono essere specificati nell'ambito della stessa operazione. L'operazione va eseguita separatamente per ciascun tipo di campo.

Visualizza campi correnti. Selezionare per visualizzare i campi per gli insiemi di dati connessi in modo attivo al nodo Anonimizza. Questa opzione è selezionata per default.

Visualizza impostazioni campi inutilizzati. Selezionare per visualizzare i campi per gli insiemi di dati non più connessi al nodo. Questa opzione è utile quando si copiano nodi da un flusso all'altro o quando si salvano e si ricaricano nodi.

Specifica del metodo di anonimizzazione dei valori di campo

La finestra di dialogo Sostituisci valori consente di scegliere se utilizzare il prefisso di default o un prefisso personalizzato per i valori di campo anonimizzati. Se si fa clic su **OK** in questa finestra di dialogo, l'impostazione di Anonimizza valori nella scheda Impostazioni viene modificata in **Sì** per il campo o i campi selezionati.

Prefisso dei valori di campo. Il prefisso di default per i valori di campo anonimizzati è **anon_**; se si preferisce utilizzare un prefisso diverso, scegliere **Personalizzato** e digitare il prefisso.

La finestra di dialogo Trasforma valori viene visualizzata solo per i campi di tipo Continuo e consente di specificare se per la trasformazione dei valori dei campi devono essere utilizzati valori casuali o valori specificati dall'utente.

Casuale. Scegliere questa opzione se si desidera utilizzare valori casuali per la trasformazione. L'opzione **Imposta seme random** è selezionata per default; specificare un valore nel campo **Seme** oppure utilizzare il valore di default.

Fisso. Scegliere questa opzione se si desidera specificare i propri valori per la trasformazione.

- **Scala per.** Il numero per il quale i valori dei campi verranno moltiplicati durante la trasformazione. Il valore minimo è 1 e il valore massimo solitamente è 10, ma può essere ridotto per evitare l'overflow.
- **Transla per.** Il numero che verrà aggiunto ai valori dei campi durante la trasformazione. Il valore minimo è 0; il valore massimo è normalmente 1000, ma può essere ridotto per evitare l'overflow.

Anonimizzazione dei valori dei campi

I campi che sono stati selezionati per l'anonimizzazione nella scheda Impostazioni subiscono l'anonimizzazione dei valori:

- Quando si esegue il flusso contenente il nodo Anonimizza
- Quando si visualizzano i valori in anteprima

Per visualizzare i valori in anteprima, fare clic sul pulsante **Anonimizza valori** nella scheda Valori anonimizzati. Selezionare quindi un nome di campo dall'elenco a discesa.

Se il livello di misurazione è Continuo, vengono visualizzati:

- Valori minimo e massimo dell'intervallo originale
- Equazione utilizzata per trasformare i valori

Se il livello di misurazione è diverso da Continuo, vengono visualizzati i valori originale e anonimizzato per tale campo.

Quando la visualizzazione presenta uno sfondo giallo, significa che l'impostazione del campo selezionato è stata modificata dopo l'ultima anonimizzazione dei valori oppure che sono state apportate delle modifiche ai dati a monte del nodo Anonimizza e i valori anonimizzati potrebbero non essere più corretti. Viene visualizzato l'insieme di valori corrente; fare clic nuovamente sul pulsante **Anonimizza valori** per generare un nuovo insieme di valori in base all'impostazione corrente.

Anonimizza valori. Consente di creare valori anonimizzati per il campo selezionato e di visualizzarli in una tabella. Se si utilizza un seme random per un campo di tipo Continuo, ogni volta che si fa clic su questo pulsante viene creato un insieme di valori diverso.

Cancella valori. Cancella i valori originali e anonimizzati dalla tabella.

Nodo Ricodifica

Il nodo Ricodifica consente la trasformazione da un insieme di valori categoriali a un altro. L'operazione di ricodifica consente di comprimere categorie o raggruppare dati per l'analisi. Per esempio, è possibile ricodificare i valori per *Prodotto* in tre gruppi, quali *Stoviglie*, *Biancheria* ed *Elettrodomestici*. Questa operazione viene spesso eseguita direttamente da un nodo distribuzione raggruppando valori e generando un nodo Ricodifica. Per ulteriori informazioni, consultare l'argomento "Utilizzo di un nodo distribuzione" a pagina 212.

La ricodifica può essere eseguita per uno o più campi simbolici. È anche possibile scegliere di sostituire i nuovi valori per il campo esistente o generare un nuovo campo.

Utilizzo di un nodo Ricodifica

Prima di utilizzare un nodo Ricodifica, valutare se un altro nodo Operazioni su campi può risultare più appropriato all'attività in corso:

- Per trasformare intervalli numerici in insiemi utilizzando un metodo automatico, come i ranghi o i percentili, è necessario utilizzare un nodo Discretizza. Per ulteriori informazioni, consultare l'argomento "Nodo Discretizza" a pagina 138.
- Per codificare manualmente intervalli numerici in insiemi, è necessario utilizzare un nodo Ricava. Se, per esempio, si desidera comprimere i valori di salario in specifiche categorie di intervalli di salari, è necessario utilizzare un nodo Ricava per definire manualmente ogni categoria.
- Per creare uno o più campi flag basati sui valori di un campo categoriale, per esempio *Tipo_mutuo*, è necessario utilizzare un nodo Crea flag.
- Per convertire un campo categoriale in campo di archiviazione numerica è possibile utilizzare un nodo Ricava. Per esempio, è possibile convertire i valori *No* e *Sì* rispettivamente in 0 e 1. Per ulteriori informazioni, consultare l'argomento "Ricodifica di valori con il nodo Ricava" a pagina 132.

Impostazione delle opzioni per il nodo Ricodifica

Per utilizzare il nodo Ricodifica è necessario seguire tre passaggi:

1. Innanzitutto, selezionare se si desidera ricodificare uno o più campi.
2. Quindi, scegliere se eseguire la ricodifica nel campo esistente o creare un nuovo campo.
3. Infine, utilizzare le opzioni dinamiche nella finestra di dialogo del nodo Ricodifica per mappare gli insiemi in base alle esigenze.

Modalità. Selezionare **Singola** per ricodificare le categorie per un campo. Selezionare **Multipla** per attivare le opzioni che consentono la trasformazione di più campi contemporaneamente.

Ricodifica in. Selezionare **Ricava** per utilizzare il campo nominale originale e derivare un campo aggiuntivo contenente i valori ricodificati. Selezionare **Campo esistente** per sovrascrivere i valori nel campo originale con le nuove classificazioni. In questo caso si tratta essenzialmente di un'operazione di riempimento.

Dopo aver specificato le opzioni di sostituzione e relative alla modalità, è necessario selezionare il campo di trasformazione e specificare i nuovi valori di classificazione utilizzando le opzioni dinamiche nella parte inferiore della finestra di dialogo. Tali opzioni variano in relazione alla modalità selezionata.

Ricodifica campo (campi). Utilizzare il pulsante Selettore di campo a destra per selezionare uno (modalità Singola) o più (modalità Multipla) campi categoriali.

Nome nuovo campo. Specificare un nome per il nuovo campo nominale contenente i valori ricodificati. Questa opzione è disponibile solo in modalità Singola se è stato selezionato **Ricava**. Se si seleziona **Campo esistente**, il nome originale del campo viene mantenuto. Se si opera in modalità Multipla, questa opzione viene sostituita con controlli per specificare l'aggiunta di un'estensione a ogni nuovo campo. Per ulteriori informazioni, consultare l'argomento "Ricodifica di più campi" a pagina 137.

Ricodifica valori. Questa tabella consente di specificare una nuova mappatura tra i valori di insieme preesistenti e quelli specificati in questo contesto.

- **Valore originale.** In questa colonna sono elencati i valori esistenti per i campi selezionati.
 - **v** Utilizzare questa colonna per digitare nuovi valori di categoria o selezionarne uno dall'elenco a discesa. Se si genera automaticamente un nodo Ricodifica utilizzando i valori di un grafico di distribuzione, tali valori verranno inseriti nell'elenco a discesa. Ciò consente di mappare rapidamente i valori esistenti su un insieme di valori noto. Per esempio, le organizzazioni sanitarie a volte raggruppano le diagnosi in modi diversi, in base alla rete o alle opzioni internazionali. Dopo una fusione o un'acquisizione, a tutte le parti verrà richiesto di ricodificare in modo uniforme i nuovi dati o i dati esistenti. Aniché reinserire manualmente tutti i valori di destinazione di un lungo elenco, è possibile leggere l'elenco principale di valori in IBM SPSS Modeler, eseguire un grafico di distribuzione per il campo *Diagnosi* e generare un nodo Ricodifica (valori) per il campo direttamente dal grafico. In questo modo tutti i valori di destinazione Diagnosi saranno disponibili nell'elenco a discesa Nuovi valori.
4. Fare clic su **Utilizza** per leggere i valori originali di uno o più campi selezionati in precedenza.
 5. Fare clic su **Copia** per incollare i valori originali nella colonna *Nuovo valore* per i campi che non sono ancora stati mappati. I valori originali non mappati verranno aggiunti all'elenco a discesa.
 6. Fare clic su **Cancella nuovo** per cancellare tutte le specifiche presenti nella colonna *Nuovo valore*. *Nota:* questa opzione non cancella i valori dall'elenco a discesa.
 7. Fare clic su **Automatico** per generare automaticamente numeri interi consecutivi per ogni valore originale. Possono essere generati solo valori interi (nessun valore reale, come 1.5, 2.5 e così via).

Per esempio, è possibile generare automaticamente numeri ID di prodotto consecutivi per i nomi di prodotto o numeri dei corsi per le offerte di un'università. Questa funzionalità corrisponde alla trasformazione di ricodifica automatica degli insiemi in IBM SPSS Statistics.

Per valori non specificati utilizza. Questa opzione consente di immettere valori non specificati nel nuovo campo. È possibile scegliere di mantenere il valore originale selezionando **Valore originale** oppure specificare un valore di default.

Ricodifica di più campi

Per mappare i valori di categoria per più campi contemporaneamente, impostare la modalità **Multipla**. In questo modo vengono attivate nuove impostazioni nella finestra di dialogo Ricodifica, descritte di seguito.

Ricodifica campi. Utilizzare il pulsante Selettore di campo a destra per selezionare i campi che si desidera trasformare. In questo modo è possibile selezionare tutti i campi contemporaneamente oppure campi di un determinato tipo, per esempio nominale o flag.

Estensione nome campo. Quando si ricodificano più campi contemporaneamente, è consigliabile specificare un'estensione comune da aggiungere a tutti i nuovi campi anziché singoli nomi di campo. Specificare un'estensione quale *ricodifica* e selezionare se inserirla come prefisso o suffisso ai nomi originali dei campi.

Classe di archiviazione e livello di misurazione per i campi ricodificati

Il nodo Ricodifica crea sempre un campo nominale dall'operazione di ricodifica. In alcuni casi, il livello di misurazione può cambiare quando si utilizza la modalità di ricodifica **Campo esistente**.

L'archiviazione del nuovo campo (il modo in cui vengono *archiviati* i dati anziché come vengono *utilizzati*) viene definita in base alle seguenti opzioni della scheda Impostazioni:

- Se nel caso di valori non specificati si sceglie di utilizzare un valore di default, il tipo di archiviazione appropriato viene determinato esaminando i nuovi valori e il valore di default. Se, per esempio, tutti i valori possono essere analizzati come numeri interi, per il campo verrà utilizzato il tipo di archiviazione come numero intero.
- Se nel caso di valori non specificati si sceglie di utilizzare i valori originali, il tipo di archiviazione viene basato sulla classe di archiviazione del campo originale. Se tutti i valori possono essere analizzati con la classe di archiviazione del campo originale, tale classe viene mantenuta; in caso contrario, la classe di archiviazione viene determinata individuando il tipo di archiviazione più appropriato che includa i valori nuovi e quelli preesistenti. Ad esempio, la ricodifica di un insieme di numeri interi { 1, 2, 3, 4, 5 } con la ricodifica 4 => 0, 5 => 0 genera un nuovo insieme di interi { 1, 2, 3, 0 }, mentre la ricodifica 4 => "Maggiore di 3", 5 => "Maggiore di 3" genera l'insieme stringa { "1", "2", "3", "Maggiore di 3" }.

Nota: se il tipo originale non era stato istanziato, verrà applicato lo stesso formato per il nuovo tipo.

Nodo Discretizza

Il nodo Discretizza consente di creare automaticamente nuovi campi nominali basati sui valori di uno o più campi continui (intervallo numerico). Per esempio, è possibile trasformare un campo continuo relativo al reddito in campo categoriale contenente gruppi di reddito di larghezza equivalente o come deviazioni dalla media. In alternativa, è possibile selezionare un campo "supervisore" categoriale per mantenere la forza dell'associazione originale tra i due campi.

La discretizzazione può essere utile per diversi motivi, fra cui:

- **Requisiti degli algoritmi.** Alcuni algoritmi, come Naive Bayes e Regressione logistica, richiedono input categoriali.
- **Prestazioni.** Alcuni algoritmi, come l'algoritmo logistico multinomiale, possono avere migliori performance se viene ridotto il numero dei valori distinti dei campi di input. Per esempio, utilizzare il valore mediano o medio per ogni bin invece dei valori originali.
- **Privacy dei dati.** Le informazioni personali riservate quali, per esempio, lo stipendio, possono essere riportate sotto forma di intervalli invece che di cifre effettive, in modo da proteggere la privacy.

Sono disponibili diversi metodi di raccolta. Dopo aver creato bin per il nuovo campo, è possibile generare un nodo Ricava basato sui punti di divisione.

Utilizzo di un nodo Discretizza

Prima di utilizzare un nodo Discretizza, valutare se un'altra tecnica potrebbe essere più appropriata per l'attività in corso:

- Per specificare manualmente punti di divisione per le categorie, come specifici intervalli di salario predefiniti, utilizzare un nodo Deriva. Per ulteriori informazioni, consultare l'argomento "Nodo Ricava" a pagina 127.
- Per creare nuove categorie per insiemi esistenti, utilizzare un nodo Ricodifica. Per ulteriori informazioni, consultare l'argomento "Nodo Ricodifica" a pagina 136.

Gestione dei valori mancanti

Il nodo Discretizza gestisce i valori mancanti come segue:

- **Vuoti definiti dall'utente.** I valori mancanti specificati come valori vuoti vengono inclusi durante la trasformazione. Ad esempio, se viene specificato -99 per indicare un valore vuoto utilizzando il nodo tipologia, tale valore verrà incluso nel processo di raccolta. Per ignorare i vuoti durante la discretizzazione, utilizzare un nodo Riempimento per sostituire i valori vuoti con il valore null di sistema.

- **Valori mancanti di sistema (\$null\$).** I valori null vengono ignorati durante la discretizzazione e restano null dopo la trasformazione.

Nella scheda Impostazioni sono incluse le opzioni per le tecniche disponibili. Nella scheda Visualizza sono riportati i punti di divisione stabiliti per i dati eseguiti precedentemente attraverso il nodo.

Impostazione delle opzioni relative al nodo Discretizza

Il nodo Discretizza consente di generare automaticamente bin (categorie) utilizzando le seguenti tecniche:

- Discretizzazione larghezza fissa
- Blocchi (conteggio o somma uguale)
- Media e deviazione standard
- Ranghi
- Ottimizzazione relativa a un campo "supervisore" categoriale

La parte inferiore della finestra cambia in modo dinamico in relazione al metodo di discretizzazione selezionato.

Campi bin. Le trasformazioni in sospenso per i campi continui (intervallo numerico) sono visualizzate in questo punto. Il nodo Discretizza consente di eseguire la discretizzazione di più campi simultaneamente. Aggiungere o rimuovere campi mediante i pulsanti a destra.

Metodo di raccolta. Selezionare il metodo utilizzato per determinare i punti di divisione per i nuovi bin di campo (categorie). Gli argomenti seguenti descrivono le opzioni disponibili in ogni caso.

Soglie bin. Specifica come vengono calcolate le soglie bin.

- **Ricalcola sempre.** I punti di divisione e le allocazioni dei bin sono sempre ricalcolati ogni volta che il nodo viene eseguito.
- **Leggi dalla scheda Valori bin se disponibile.** I punti di divisione e le allocazioni dei bin sono ricalcolati solo se necessario (ad esempio, quando vengono aggiunti nuovi dati).

I metodi di discretizzazione disponibili sono esaminati negli argomenti seguenti.

Bin a larghezza fissa

Quando si sceglie **Larghezza fissa** come metodo di discretizzazione, viene visualizzato un nuovo insieme nella finestra di dialogo.

Estensione nome. Specificare un'estensione da utilizzare per i campi generati. *_BIN* è l'estensione di default. È inoltre possibile specificare se l'estensione verrà aggiunta all'inizio (**Prefisso**) o alla fine (**Suffisso**) del nome del campo. Per esempio, è possibile generare un nuovo campo denominato *reddito_BIN*.

Larghezza bin. Specificare un valore (intero o reale) utilizzato per calcolare la "larghezza" del bin. Per esempio, è possibile utilizzare il valore di default, 10, per discretizzare il campo *Età*. Poiché il campo *Età* ha un intervallo 18–65, i bin generati saranno quelli illustrati nella seguente tabella.

Tabella 19. Bin per Età con intervallo 18–65

Bin 1	Bin 2	Bin 3	Bin 4	Bin 5	Bin 6
>=da 13 a <23	>=da 23 a <33	>=da 33 a <43	>=da 43 a <53	>=da 53 a <63	>=da 63 a <73

L'inizio dell'intervallo dei bin viene calcolato sottraendo la metà della larghezza del bin al valore più basso rilevato (come indicato). Ad esempio, nei bin mostrati in precedenza, 13 viene utilizzato per iniziare gli intervalli in base al seguente calcolo: $18 [\text{valore dati più basso}] - 5 [0.5 \times (\text{larghezza bin } 10)] = 13$.

N. di bin. Questa opzione consente di specificare un numero intero utilizzato per determinare il numero di bin a larghezza fissa (categorie) per i nuovi campi.

Dopo aver eseguito il nodo Discretizza in un flusso, è possibile visualizzare le soglie dei bin generate facendo clic sulla scheda **Anteprima** nella finestra di dialogo del nodo Discretizza. Per ulteriori informazioni, consultare l'argomento "Anteprima dei bin generati" a pagina 143.

Blocchi (conteggio o somma uguale)

Il metodo di discretizzazione a blocchi crea dei campi nominali che possono essere utilizzati per suddividere i record analizzati in gruppi percentili (o quartili, decili e così via) in modo che ogni gruppo contenga lo stesso numero di record o che la somma dei valori di ciascun gruppo sia uguale. I record vengono classificati in ordine crescente a seconda del valore del campo bin specificato, in modo tale che ai record con i valori più bassi per la variabile bin selezionata viene assegnato il rango 1, all'insieme di record seguente il rango 2 e così via. I valori di soglia di ogni bin vengono generati automaticamente in base ai dati e al metodo di discretizzazione a blocchi utilizzato.

Estensione nome blocco. Specificare un'estensione da utilizzare per i campi generati con p-tili standard. L'estensione di default è `_TILE` preceduta da *N*, dove *N* è il numero percentile. È inoltre possibile specificare se l'estensione verrà aggiunta all'inizio (**Prefisso**) o alla fine (**Suffisso**) del nome del campo. Per esempio, è possibile generare un nuovo campo denominato `reddito_BIN4`.

Estensione blocco personalizzata. Specificare un'estensione da utilizzare per l'intervallo di blocchi personalizzato. L'impostazione di default è `_TILEN`. Notare che *N*, in questo caso, *non* viene sostituito dal numero personalizzato.

I p-tili disponibili sono:

- **Quartile.** Genera quattro bin, ognuno contenente il 25% dei casi.
- **Quintile.** Genera cinque bin, ognuno contenente il 20% dei casi.
- **Decile.** Genera dieci bin, ognuno contenente il 10% dei casi.
- **Ventile.** Genera venti bin, ognuno contenente il 5% dei casi.
- **Percentile.** Genera cento bin, ognuno contenente l'1% dei casi.
- **N personalizzato.** Selezionare per specificare il numero di bin. Per esempio, il valore 3 produrrebbe tre categorie suddivise (due punti di divisione), ognuna contenente il 33.3% dei casi.

Si noti che se nei dati è presente un numero di valori discreti inferiore al numero di blocchi specificati, non tutti i blocchi saranno utilizzati. In tali casi, la nuova distribuzione risulterà simile a quella originale dei dati.

Metodo di espansione. Consente di specificare il metodo utilizzato per assegnare i record ai bin.

- **Conteggio dei record.** Cerca di assegnare un numero di record uguale a ogni bin.
- **Somma dei valori.** Cerca di assegnare i record ai bin in modo che la somma dei valori di ogni bin sia uguale. Per esempio, nella targetizzazione dei tentativi di vendita, questo metodo può essere utilizzato per assegnare i potenziali clienti a gruppi di decili in base al valore di ogni record, con i potenziali clienti di valore massimo nel primo bin. Per esempio un'azienda farmaceutica potrebbe classificare i medici in gruppi di decili in base al numero di ricette compilate. Ogni decile conterrebbe più o meno lo stesso numero di ricette, ma il numero di individui che genera le ricette non sarebbe uguale, poiché gli individui che compilano più ricette sono concentrati nel decile 10. Questo approccio presuppone che tutti i valori siano maggiori di zero e in caso contrario può generare risultati inattesi.

Pari merito. Se i valori a entrambi i lati di un punto di divisione sono identici, si ottiene una condizione di pari merito. Per esempio, nel caso di assegnazione di decili in cui più del 10% dei record ha lo stesso valore per il campo bin, non è possibile che tutti i record possano rientrare nello stesso bin senza forzare in qualche modo la soglia. I record a pari merito possono essere spostati al bin successivo o mantenuti in

quello corrente, ma è necessario che la condizione venga risolta affinché tutti i record con valori identici rientrino nello stesso bin, anche se ciò potrebbe comportare per alcuni bin un numero di record superiore al previsto. Di conseguenza, è possibile che anche le soglie dei bin successivi vengano adeguate e che i valori siano assegnati in modo diverso per lo stesso insieme di numeri in base al metodo usato per risolvere le condizioni di pari merito.

- **Aggiungi a successivo.** Selezionare questa opzione per spostare i valori pari merito nel bin successivo.
- **Mantieni in corrente.** Mantiene i valori pari merito nel bin corrente (più basso). Questo metodo può determinare la creazione di un numero totale di bin inferiore.
- **Assegna casualmente.** Selezionare questa opzione per allocare in modo casuale i valori pari merito a un bin. In questo modo si tenta di mantenere uguale il numero di record di ogni bin.

Esempio: espansione per conteggio dei record

La tabella riportata di seguito illustra il modo in cui i valori di campo semplificati vengono classificati come quartili quando viene eseguita l'espansione per conteggio di record. Si noti che i risultati possono variare a seconda dell'opzione di pari merito selezionata.

Tabella 20. Esempio di espansione per conteggio di record.

Valori	Aggiungi a successivo	Mantieni in corrente
10	1	1
13	2	1
15	3	2
15	3	2
20	4	3

Il numero di elementi per bin viene calcolato come segue:

numero totale del valore/numero di blocchi

Nel precedente esempio semplificato il numero desiderato di elementi per bin è 1,25 (5 valori/4 quartili). Il valore 13 (ossia il valore numero 2) si trova a cavallo della soglia di conteggio desiderata 1,25 e, quindi, viene trattato diversamente in relazione all'opzione di pari merito selezionata. In modalità **Aggiungi a successivo** questo valore viene aggiunto al bin 2, mentre in modalità **Mantieni in corrente** viene lasciato nel bin 1, spostando l'intervallo di valori del bin 4 all'esterno dei valori di dati esistenti. Come risultato, vengono creati solo tre bin e le soglie per ciascun bin vengono regolate di conseguenza, come illustrato nella tabella riportata di seguito.

Tabella 21. Risultato esempio di raccolta.

Intervallo	Inferiore	Superiore
1	>=10	<15
2	>=15	<20
3	>=20	<=20

Nota: l'attivazione dell'elaborazione parallela può accelerare il processo di discretizzazione a blocchi.

Classifica casi

Se si sceglie **Ranghi** come metodo di discretizzazione, nella finestra di dialogo viene visualizzato un nuovo insieme di opzioni.

L'assegnazione di ranghi crea nuovi campi contenenti ranghi, ranghi frazionari e valori percentili per i campi numerici in base alle opzioni specificate di seguito.

Ordine rango. Selezionare **Crescente** (il valore più basso viene indicato con 1) o **Decrescente** (il valore più alto viene indicato con 1).

Rango. Selezionare il tipo di rango in ordine crescente o decrescente, come indicato in precedenza. L'intervallo dei valori nel nuovo campo sarà $1-N$, dove N è il numero di valori discreti nel campo originale. Ai valori pari merito viene assegnata la media del rispettivo rango.

Rango frazionario. Selezionare i tipi di rango in cui il valore del nuovo campo equivale al rango diviso per la somma dei pesi dei casi non mancanti. I ranghi frazionari sono compresi nell'intervallo 0–1.

Percentuale rango frazionario. Ciascun rango è diviso per il numero di record con valori validi e moltiplicato per 100. I ranghi frazionari percentuali sono compresi nell'intervallo 1–100.

Estensione. Per tutte le opzioni dei ranghi è possibile creare estensioni personalizzate e specificare se l'estensione verrà aggiunta all'inizio (**Prefisso**) o alla fine (**Suffisso**) del nome del campo. Per esempio, è possibile generare un nuovo campo denominato *reddito_P_RANGO*.

Deviazione standard/Media

Se si sceglie **Deviazione standard/Media** come metodo di discretizzazione, nella finestra di dialogo viene visualizzato un nuovo insieme di opzioni.

Questo metodo genera uno o più campi nuovi con categorie suddivise in base ai valori della deviazione standard e media della distribuzione dei campi specificati. Selezionare il numero di deviazioni da utilizzare di seguito.

Estensione nome. Specificare un'estensione da utilizzare per i campi generati. *_SDBIN* è l'estensione di default. È inoltre possibile specificare se l'estensione verrà aggiunta all'inizio (**Prefisso**) o alla fine (**Suffisso**) del nome del campo. Per esempio, è possibile generare un nuovo campo denominato *reddito_SDBIN*.

- **+/- 1 deviazione standard.** Selezionare questa opzione per generare tre bin.
- **+/- 2 deviazioni standard.** Selezionare questa opzione per generare cinque bin.
- **+/- 3 deviazioni standard.** Selezionare questa opzione per generare sette bin.

Ad esempio, se si seleziona +/-1 deviazione standard, vengono calcolati i tre bin illustrati nella tabella riportata di seguito.

Tabella 22. Esempio di bin con deviazione standard.

Bin 1	Bin 2	Bin 3
$x < (\text{Media} - \text{Dev std.})$	$(\text{Media} - \text{Dev std.}) \leq x \leq (\text{Media} + \text{Dev std.})$	$x > (\text{Dev Std.} + \text{Media})$

In una distribuzione normale, il 68% dei casi è compreso in una deviazione standard della media, il 95% in due deviazioni standard e il 99% in tre deviazioni standard. Tuttavia, si tenga presente che la creazione di categorie suddivise in base a deviazioni standard può determinare la definizione di alcuni bin all'esterno dell'effettivo intervallo di dati e anche dell'intervallo di possibili valori di dati (per esempio, un intervallo di salari negativo).

Categorizzazione ottimale

Se il campo da discretizzare è fortemente associato a un altro campo categoriale, è possibile selezionare il campo categoriale come campo "supervisore" al fine di creare i bin in modo che venga mantenuta la forza dell'associazione originale tra i due campi.

Per esempio, si supponga di avere utilizzato l'analisi dei cluster per raggruppare diversi stati o regioni in base ai tassi di insolvenza sui mutui residenziali, con i tassi più alti raccolti nel primo cluster. In questo caso, si potrebbe scegliere *Percentuale oltre la scadenza* e *Percentuale pignoramenti* come campi bin e il campo di appartenenza al cluster generato dal modello come campo supervisore.

Estensione nome. Specificare un'estensione da utilizzare per il campo o i campi generati e indicare se si desidera aggiungerla all'inizio (**Prefisso**) o alla fine (**Suffisso**) del nome del campo. Per esempio, è possibile generare un nuovo campo denominato *scaduti_OTTIMALE* e un altro campo denominato *pignorati_OTTIMALE*.

Campo supervisore. Un campo categoriale utilizzato per creare i bin.

Esegui pre-discretizzazione per ottimizzare le performance con grandi insiemi di dati. Indica se è necessario utilizzare un'elaborazione preliminare per facilitare la discretizzazione ottimale. Il metodo raggruppa i valori di scala in un alto numero di bin utilizzando un semplice metodo di discretizzazione non supervisionato, rappresenta i valori dei singoli bin in base alla media e corregge di conseguenza il peso del caso prima di procedere alla discretizzazione supervisionata. In pratica, questo metodo privilegia la velocità rispetto alla precisione ed è consigliato per insiemi di dati di grosse dimensioni. È anche possibile specificare il numero massimo di bin di ogni variabile dopo la pre-elaborazione quando si utilizza questa opzione.

Unisci bin che presentano conteggi relativamente bassi con un elemento adiacente di dimensioni maggiori. Se attivata, questa opzione indica che un bin viene unito se il rapporto tra le sue dimensioni (numero di casi) e quelle di un bin adiacente è inferiore alla soglia specificata; si noti che soglie più elevate possono determinare un maggior numero di unioni.

Impostazioni dei punti di divisione

La finestra di dialogo Impostazioni dei punti di divisione consente di specificare le opzioni avanzate relative all'algoritmo di discretizzazione ottimale. Queste opzioni indicano all'algoritmo in che modo calcolare i bin utilizzando il campo obiettivo.

Discretizza punti di fine. Consente di specificare se i punti di fine inferiore e superiore devono essere inclusi ($\text{inferiore} \leq x$) o esclusi ($\text{inferiore} < x$).

Primo e ultimo bin. È possibile specificare se il primo e l'ultimo bin devono essere illimitati (estendersi verso l'infinito positivo o negativo) oppure limitati dai punti dati inferiore o superiore.

Anteprima dei bin generati

La scheda Discretizza valori del nodo Discretizza consente di visualizzare le soglie per i bin generati. Mediante il menu Genera, è anche possibile creare un nodo Deriva che può essere utilizzato per applicare queste soglie da un dataset all'altro.

Campo nei bin. Utilizzare l'elenco a discesa per selezionare un campo da visualizzare. I nomi di campo mostrati utilizzano il nome originale del campo per maggiore chiarezza.

Blocco. Utilizzare l'elenco a discesa per selezionare un blocco, per esempio 10 o 100, da visualizzare. Questa opzione è disponibile solo quando i bin sono stati generati utilizzando il metodo dei blocchi con conteggio o somma uguale.

Soglie bin. Vengono qui visualizzati i valori di soglia per ogni bin generato, nonché il numero dei record che rientrano in ogni bin. Solo nel caso del metodo di discretizzazione ottimale, il numero dei record dei singoli bin viene espresso come percentuale della quantità totale. Si noti che le soglie non sono applicabili quando si utilizza il metodo di discretizzazione per ranghi.

Leggi valori. Consente di leggere i valori discretizzati dall'insieme di dati. Si noti che anche le soglie vengono sovrascritte quando nuovi dati vengono eseguiti attraverso il flusso.

Generazione di un nodo Deriva

Il menu Genera consente di creare un nodo Deriva sulla base delle soglie correnti. È utile per applicare soglie di bin stabilite da un dataset all'altro. Inoltre, dopo aver individuato i punti di suddivisione, un'operazione Deriva risulta più efficace (più rapida) di un'operazione Discretizza quando si utilizzano insiemi di dati molto estesi.

Nodo Analisi RFM

Il nodo Analisi RFM (Recency, Frequency, Monetary, Passato recente, Frequenza, Monetario) consente di determinare in modo quantitativo i clienti potenzialmente migliori verificando quanto tempo è trascorso dal loro ultimo acquisto (passato recente), con quale frequenza hanno effettuato acquisti (frequenza) e quanto hanno speso per tutte le transazioni (monetario).

L'analisi RFM si basa sull'idea che i clienti che acquistano una volta un prodotto o un servizio hanno maggiori probabilità di acquistarlo nuovamente. I dati sui clienti suddivisi in categorie vengono distribuiti in un certo numero di bin, con i criteri di discretizzazione modificati in funzione delle proprie esigenze. In ogni bin, ai clienti viene assegnato un punteggio; i punteggi vengono quindi combinati per ottenere un punteggio RFM complessivo. Tale punteggio è una rappresentazione dell'appartenenza di un cliente ai bin creati per ogni parametro RFM. Questi dati discretizzati possono essere sufficienti per le proprie esigenze, per esempio perché individuano i clienti di maggior valore e che acquistano con maggiore frequenza; in caso contrario, i dati possono essere passati a un flusso per essere sottoposti a ulteriore modellazione e analisi.

Si noti tuttavia che, sebbene la capacità di analizzare e classificare i punteggi RFM sia uno strumento utile, è necessario tenere presente determinati fattori quando lo si utilizza. Si può tentare di scegliere come obiettivo i clienti con le classificazioni più elevate; tuttavia, una eccessiva sollecitazione di questi clienti potrebbe creare fastidio e determinare invece un calo degli acquisti ripetuti. È inoltre bene ricordare che i clienti con punteggio basso non vanno trascurati, bensì coltivati, per farli diventare clienti migliori. Al contrario, a seconda del mercato, un punteggio alto in sé non indica necessariamente un buon cliente potenziale. Per esempio, un cliente allocato al bin 5 per passato recente (cioè un cliente che ha effettuato un acquisto molto di recente) potrebbe non essere il cliente obiettivo migliore per un'azienda che vende prodotti costosi e di maggiore durata quali auto o televisori.

Nota: in base al modo in cui i dati sono archiviati, potrebbe essere necessario far precedere un nodo Aggregazione RFM al nodo Analisi RFM per trasformare i dati in un formato utilizzabile. Per esempio, i dati di input devono essere nel formato del cliente, con una riga per ogni cliente; se i dati dei clienti sono in formato transazionale, utilizzare un nodo Aggregazione RFM a monte per derivare i campi Passato recente, Frequenza e Monetario. Per ulteriori informazioni, consultare l'argomento "Nodo Aggregazione RFM" a pagina 72.

I nodi Aggregazione RFM e Analisi RFM di IBM SPSS Modeler sono impostati per utilizzare la discretizzazione indipendente, ovvero classificano e discretizzano i dati rispetto a ogni misura di passato recente, frequenza e valore monetario, indipendentemente dai loro valori o dalle altre due misure.

Impostazioni del nodo Analisi RFM

Passato recente. Il Selettore di campo (campo a destra della casella di testo) permette di selezionare il campo Passato recente il cui valore può essere una data, un timestamp o un semplice numero. Si noti che, quando la data della transazione più recente è rappresentata da una data o da un timestamp, si considera più recente il valore più alto; quando invece è rappresentata da un numero, questo indica il tempo trascorso dalla transazione più recente e si considera più recente il valore più basso.

Nota: se nel flusso il nodo Analisi RFM è preceduto da un nodo Aggregazione RFM, i campi Passato recente, Frequenza e Monetario generati dal nodo Aggregazione RFM devono essere selezionati come input nel nodo Analisi RFM.

Frequenza. Selezionare il campo Frequenza da utilizzare con il selettore di campo.

Monetario. Selezionare il campo Monetario da utilizzare con il Selettore di campo.

Numero di bin. Selezionare il numero di bin da creare per ciascuno dei tre tipi di output. Il valore predefinito è 5.

Nota: il numero minimo di bin è 2, il massimo è 9.

Peso. Per default, nel calcolo dei punteggi la massima importanza viene attribuita ai dati relativi al passato recente, seguiti da quelli di frequenza e monetario. Se necessario è possibile modificare la ponderazione che influisce su uno o più di questi dati per cambiare il dato a cui attribuire la massima importanza.

Il punteggio RFM viene calcolato nel modo seguente: (punteggio Passato recente x peso Passato recente) + (punteggio Frequenza x peso Frequenza) + (punteggio Monetario x peso Monetario).

Pari merito. Specifica le modalità di discretizzazione dei punteggi identici (pari merito). Le opzioni disponibili sono:

- **Aggiungi a successivo.** Selezionare questa opzione per spostare i valori pari merito nel bin successivo.
- **Mantieni in corrente.** Mantiene i valori pari merito nel bin corrente (più basso). Questo metodo può determinare la creazione di un numero totale di bin inferiore. Questa è l'impostazione di default.

Soglie bin. Specificare se i punteggi RFM e le allocazioni dei bin vengono sempre rielaborati all'esecuzione del nodo, o se vengono elaborati solo quando è necessario (per esempio in occasione dell'aggiunta di nuovi dati). Se si seleziona **Leggi dalla scheda Valori bin se disponibile**, è possibile modificare i punti di divisione superiore ed inferiore per i diversi bin nella scheda Discretizza valori.

Quando viene eseguito, il nodo Analisi RFM discretizza i campi grezzi di passato recente, frequenza e monetario e aggiunge i seguenti nuovi campi all'insieme di dati:

- Punteggio attualità. Un rango (valore di discretizzazione) per Passato recente
- Punteggio frequenza. Un rango (valore di discretizzazione) per la Frequenza
- Punteggio valore monetario. Un rango (valore di discretizzazione) per Monetario
- Punteggio RFM. La somma ponderata dei punteggi di passato recente, frequenza e monetario.

Aggiungi valori anomali ai bin alle estremità. Se viene selezionata questa casella di controllo, i record che si trovano al di sotto del bin inferiore vengono aggiunti al bin inferiore ed i record che si trovano al di sopra del bin superiore vengono aggiunti al bin superiore — in caso contrario, viene loro assegnato un valore null. La casella è disponibile solo se si seleziona **Leggi dalla scheda Valori bin se disponibile**.

Discretizzazione con il nodo Analisi RFM

La scheda Discretizza valori consente di visualizzare e in alcuni casi modificare le soglie per i bin generati.

Nota: i valori di questa scheda si possono modificare solo se si è selezionata l'opzione **Leggi dalla scheda Valori bin se disponibile** nella scheda Impostazioni.

Campo nei bin. Utilizzare l'elenco a discesa per selezionare un campo da suddividere in bin. I valori disponibili sono quelli selezionati nella scheda Impostazioni.

Tabella Discretizza valori. I valori delle soglie vengono visualizzati per ogni bin generato. Se si seleziona **Leggi dalla scheda Valori bin se disponibile** nella scheda Impostazioni è possibile modificare i punti di divisione superiore e inferiore di ogni bin facendo doppio clic sulla cella corrispondente.

Leggi valori. Consente di leggere i valori discretizzati dall'insieme di dati e di popolare la tabella Discretizza valori. Si noti che se si seleziona **Ricalcola sempre** nella scheda Impostazioni, le soglie dei bin verranno sovrascritte quando vengono elaborati nuovi dati nel flusso.

Nodo Partizione

I nodi Partizione vengono utilizzati per generare un campo partizione che suddivide i dati in sottoinsiemi distinti o campioni destinati alle fasi di addestramento, test e convalida della creazione del modello. Utilizzando un campione per generare il modello e un altro campione per sottoporlo a verifica, è possibile ottenere una valida indicazione del modo in cui il modello potrà essere esteso a insiemi di dati di dimensioni maggiori, più simili ai dati correnti.

Il nodo Partizione genera un campo nominale con il ruolo impostato su **Partizione**. In alternativa, se nei dati esiste già un campo appropriato, è possibile definirlo come una partizione utilizzando un nodo Tipo. In questo caso non è necessario un nodo Partizione distinto. Come partizione è possibile utilizzare qualsiasi campo nominale istanziato con due o tre valori, ma non è possibile utilizzare i campi flag. Per ulteriori informazioni, consultare l'argomento "Impostazione del ruolo del campo" a pagina 119.

È possibile definire più campi partizione in un flusso, ma in questo caso nella scheda Campi di ogni nodo Modelli che utilizza il partizionamento è necessario selezionare un singolo campo partizione. Se è presente un'unica partizione, verrà utilizzata automaticamente quando si attiva il partizionamento.

Attivazione del partizionamento. Per utilizzare la partizione in un'analisi, è necessario attivare il partizionamento nella scheda Opzioni modello del nodo di creazione del modello o di analisi appropriato. Se si deselecta questa opzione, è possibile disattivare il partizionamento senza rimuovere il campo.

Per creare un campo partizione in base ad altri criteri, per esempio un intervallo di date o una posizione, è inoltre possibile utilizzare un nodo Ricava. Per ulteriori informazioni, consultare l'argomento "Nodo Ricava" a pagina 127.

Esempio. Durante la creazione di un flusso RFM per individuare i clienti recenti che hanno risposto in modo favorevole alle precedenti campagne di marketing, la divisione marketing di un'azienda utilizza un nodo Partizione per suddividere i dati in partizioni di addestramento e di test.

Opzioni del nodo Partizione

Campo partizione. Specifica il nome del campo creato dal nodo.

Partizioni. È possibile partizionare i dati in due campioni (addestramento e test) o in tre campioni (addestramento, test e convalida).

- **Addestramento e test.** Partiziona i dati in due campioni, consentendo l'addestramento del modello con un campione e il test con un altro.
- **Addestramento, test e convalida.** Partiziona i dati in tre campioni, consentendo l'addestramento del modello con un campione, il test e il perfezionamento del modello con un secondo campione e la convalida dei risultati con un terzo. In questo modo, tuttavia, le dimensioni di ogni partizione vengono ridotte di conseguenza e ciò risulta particolarmente utile se si utilizzano insiemi di dati di dimensioni molto grandi.

Dimensione delle partizioni. Specifica la dimensione relativa di ogni partizione. Se la somma delle dimensioni di partizione è inferiore al 100%, i record non inclusi in una partizione verranno scartati. Per esempio, se un utente ha 10 milioni di record e ha specificato che il 5% delle dimensioni delle partizioni sono di addestramento e il 10% sono di test, dopo avere eseguito il nodo saranno presenti approssimativamente 500.000 record di addestramento e un milione di record di test, mentre i rimanenti saranno scartati.

Valori. Specifica i valori utilizzati per rappresentare ogni campione di partizione nei dati.

- **Utilizza valori definiti dal sistema ("1," "2" e "3").** Utilizza un intero per rappresentare ogni partizione. Per esempio, tutti i record che fanno parte del campione di addestramento hanno un valore 1 per il campo partizione. Ciò garantisce che i dati saranno portabili con opzioni internazionali diverse e che il criterio di ordinamento verrà mantenuto, in modo tale che 1 continuerà a rappresentare la partizione di addestramento, se il campo partizione viene reistanziato in una posizione diversa, per esempio per la lettura di dati da un database. I valori richiedono tuttavia un'interpretazione.
- **Accoda etichette a valori definiti dal sistema.** Combina l'intero con un'etichetta, per esempio i record della partizione di addestramento hanno il valore *1_Addestramento*. In questo modo, un utente che osserva i dati è in grado di identificare i valori e il criterio di ordinamento viene mantenuto. I valori sono tuttavia specifici delle opzioni internazionali impostate.
- **Utilizza etichette come valori.** Utilizza l'etichetta senza interi, per esempio *Addestramento*. Ciò consente di definire i valori mediante la modifica delle etichette. I dati sono tuttavia correlati alle opzioni internazionali specifiche e la nuova creazione istanza di una colonna della partizione determinerà il loro inserimento nel relativo criterio di ordinamento naturale, che potrebbe corrispondere al relativo ordine "semantico".

Imposta seed random. Quando si esegue il campionamento o il partizionamento di record in base a una percentuale casuale, questa opzione consente di duplicare gli stessi risultati in un'altra sessione. Specificando il valore iniziale utilizzato dal generatore di numeri random, è possibile garantire che vengano assegnati gli stessi record a ogni esecuzione del nodo. Immettere il valore seme desiderato, oppure fare clic sul pulsante **Genera** per generare automaticamente un valore casuale. Se questa opzione non è selezionata, verrà generato un campione diverso ogni volta che si esegue il nodo.

Nota: quando si utilizza l'opzione **Imposta seme random** con record letti da un database, potrebbe essere necessario un nodo Ordina prima di eseguire il campionamento, per garantire lo stesso risultato ogni volta che viene eseguito il nodo. Questo si verifica perché il seme random dipende dall'ordine dei record, per il quale non si ha la garanzia che rimanga invariato in un database relazionale. Per ulteriori informazioni, consultare l'argomento "Nodo Ordina" a pagina 73.

Consente a SQL di assegnare record alle partizioni. (Solo per database di Livello 1) Selezionare questa casella se si desidera utilizzare il push back SQL per assegnare record alle partizioni. Dall'elenco a discesa **Campo unico**, scegliere un campo con valori univoci (come un campo ID) per fare in modo che i record vengano assegnati in modo casuale ma ripetibile.

I livelli di database sono spiegati nella descrizione del nodo origine del database. Per ulteriori informazioni, consultare l'argomento "Nodo origine del database" a pagina 13.

Generazione di nodi Seleziona

Mediante il menu Genera del nodo Partizione è possibile generare automaticamente un nodo Seleziona per ogni partizione. Per esempio, è possibile selezionare tutti i record della partizione di addestramento per ottenere ulteriori valutazioni o analisi utilizzando solo questa partizione.

Nodo Crea flag

Il nodo Crea flag consente di creare campi flag in base ai valori categoriali di uno o più campi nominali. Per esempio, l'insieme di dati può contenere un campo nominale, *PRESSIONE*, con i valori *Alto*, *Normale* e *Basso*. Per facilitare la manipolazione dei dati, è possibile creare un campo flag per la pressione alta, che indica se il paziente ha una pressione sanguigna elevata oppure no.

Impostazione delle opzioni relative al nodo Crea flag

Imposta campi. Elenca tutti i campi di dati con livello di misurazione *Nominale* (insieme). Selezionarne uno dall'elenco per visualizzare i valori nell'insieme. È possibile scegliere tra questi valori per creare un campo flag. Si noti che per poter visualizzare i campi nominali disponibili (e i relativi valori) è necessario

che i dati vengano completamente istanziati mediante un nodo origine o Tipo a monte. Per ulteriori informazioni, consultare l'argomento "Nodo Tipo" a pagina 111.

Estensione nome campo. Selezionare questa opzione per attivare i controlli per la definizione di un'estensione che verrà aggiunta al nuovo campo flag come suffisso o come prefisso. Per default, i nuovi nomi di campo vengono creati automaticamente, mediante la combinazione del nome originale del campo con il valore del campo in un'etichetta, per esempio *Nomecampo_valorecampo*.

Valori disponibili. I valori nell'insieme selezionato sopra vengono visualizzati in questa casella. Selezionare uno o più valori per i quali si desidera generare i flag. Per esempio, se i valori in un campo denominato *pressione_sanguigna* sono *Alto*, *Normale* e *Basso*, è possibile selezionare *Alto* e aggiungere questo valore all'elenco a destra. In questo modo verrà creato un campo con un flag per i record contenenti un valore che indica un valore elevato di pressione sanguigna.

Crea campi flag. I campi flag creati vengono elencati in questa casella. È possibile specificare opzioni per l'assegnazione di un nome al nuovo campo utilizzando i controlli dell'estensione dei nomi di campo.

Valore vero. Specificare il valore vero utilizzato dal nodo durante l'impostazione di un flag. Per default, questo valore è T.

Valore falso. Specificare il valore falso utilizzato dal nodo durante l'impostazione di un flag. Per default, questo valore è F.

Chiavi di aggregazione. Selezionare questa opzione per raggruppare i record in base ai campi chiave specificati sotto. Se si seleziona **Chiavi di aggregazione**, tutti i campi flag in un gruppo verranno "attivati" se *qualsiasi* record è stato impostato su vero. Utilizzare il Selettore di campo per specificare i campi chiave che verranno utilizzati per l'aggregazione dei record.

Nodo Riorganizza

Il nodo Riorganizza può essere utilizzato per generare più campi in base ai valori di un campo nominale o flag. I campi così generati possono contenere valori di altri campi o flag numerici (0 e 1). Il funzionamento di questo nodo è simile a quello del nodo Crea flag, ma offre maggiore flessibilità. Esso consente, infatti, di creare campi di qualunque tipo (anche flag numerici) utilizzando i valori di un altro campo. In seguito è possibile eseguire l'aggregazione o altri tipi di manipolazione con altri nodi downstream. Il nodo Crea flag consente di aggregare i campi con un'unica operazione, il che può risultare comodo in caso di creazione di campi flag.

Per esempio, l'insieme di dati che segue contiene un campo nominale, *Conto*, con i valori *Risparmio* e *Bozza*. Per ogni conto vengono registrati il saldo d'apertura e il saldo attuale e ad alcuni clienti sono intestati più conti di ciascun tipo. Si supponga di voler sapere se ogni cliente ha un determinato tipo di conto e, in caso affermativo, quanto denaro è presente in ogni tipo di conto. Con il nodo Riorganizza, viene generato un campo per ogni singolo valore del *Conto* e si seleziona *Saldo_attuale* come valore. Ogni nuovo campo conterrà il saldo attuale relativo al record corrispondente.

Tabella 23. Dati di esempio prima della riorganizzazione.

IDclie	Conto	Saldo_apert	Saldo_att
12701	Bozza	1000	1005.32
12702	Risparmio	100	144.51
12703	Risparmio	300	321.20
12703	Risparmio	150	204.51
12703	Bozza	1200	586.32

Tabella 24. Dati di esempio dopo la riorganizzazione.

IDclie	Conto	Saldo_apert	Saldo_att	Conto_Bozza_Saldo_att	Conto_risparmio_Saldo_att
12701	Bozza	1000	1005.32	1005.32	\$null\$
12702	Risparmio	100	144.51	\$null\$	144.51
12703	Risparmio	300	321.20	\$null\$	321.20
12703	Risparmio	150	204.51	\$null\$	204.51
12703	Bozza	1200	586.32	586.32	\$null\$

Utilizzo del nodo Riorganizza con il nodo Aggregazione

In molti casi può essere consigliabile abbinare il nodo Riorganizza a un nodo Aggregazione. Nell'esempio precedente, un cliente (con il codice identificativo 12703) è titolare di tre conti. È possibile utilizzare un nodo Aggregazione per calcolare il saldo totale di ogni tipo di conto. Il campo chiave è *IDclie* e i campi aggregati sono i nuovi campi riorganizzati, *Conto_Bozza_Saldo_att* e *Conto_Risparmio_Saldo_att*. I risultati sono illustrati nella tabella seguente.

Tabella 25. Dati di esempio dopo la riorganizzazione e l'aggregazione.

IDclie	Conteggio_record	Conto_Bozza_Somma_Saldo_att	Conto_Risparmio_Somma_Saldo_att
12701	1	1005.32	\$null\$
12702	1	\$null\$	144.51
12703	3	586.32	525.71

Impostazione delle opzioni relative al nodo Riorganizza

Campi disponibili. Elenca tutti i campi di dati con livello di misurazione *Nominale* (insieme) o *Flag*. Selezionare un campo per visualizzare i valori dell'insieme o del flag, quindi scegliere i valori per la creazione dei campi riorganizzati. Si noti che per poter visualizzare i campi disponibili (e i relativi valori), è necessario che i dati vengano completamente istanziati mediante un nodo origine o Tipo a monte. Per ulteriori informazioni, consultare l'argomento "Nodo Tipo" a pagina 111.

Valori disponibili. I valori nell'insieme selezionato sopra vengono visualizzati in questa casella. Selezionare uno o più valori per i quali si desidera generare i campi riorganizzati. Per esempio, se i valori in un campo denominato *Pressione sanguigna* sono *Alto*, *Normale* e *Basso*, è possibile selezionare *Alto* e aggiungere questo valore all'elenco a destra. Questo determinerà la creazione di un campo con il valore specificato (vedere sotto) per i record con valore *Alto*.

Crea campi riorganizzati. I campi riorganizzati creati vengono elencati in questa casella. Per default, i nuovi nomi di campo vengono creati automaticamente, mediante la combinazione del nome originale del campo con il valore del campo in un'etichetta, per esempio *Nomecampo_valorecampo*.

Includi nomi campi. Deselezionare questa opzione per rimuovere il nome del campo originale apposto come prefisso dai nomi dei nuovi campi.

Utilizza valori da altri campi. Consente di specificare uno o più campi i cui valori saranno utilizzati per popolare i campi riorganizzati. Utilizzare il Selettore di campo per selezionare uno o più campi. Per ogni campo scelto viene creato un nuovo campo. Il nome del campo del valore viene accodato al nome del campo riorganizzato — ad esempio, *Pressione_Alto_Età* o *Pressione_Basso_Età*. Ogni nuovo campo eredita il tipo di campo del valore originale.

Crea flag di valore numerici. Selezionare questa opzione per popolare i nuovi campi con flag di valore numerici (0 = falso e 1 = vero) anziché utilizzare un valore preso da un altro campo.

Nodo Trasponi

Per default, le colonne sono campi e le righe sono record o osservazioni. Se necessario, è possibile utilizzare un nodo Trasponi per scambiare i dati delle righe e delle colonne in modo da trasformare i campi in record e i record in campi. Per esempio, se nei dati di serie temporali di cui si dispone ogni serie è una riga anziché una colonna, è possibile trasporre i dati prima di analizzarli.

Impostazione delle opzioni del nodo Trasponi

Nomi di campo nuovi

I nomi di nuovi campi possono essere generati automaticamente in base ad un prefisso specificato oppure letti da un campo esistente nei dati.

Utilizza prefisso. Questa opzione genera automaticamente i nomi dei nuovi campi in base al prefisso specificato (*Campo1*, *Campo2* e così via). Il prefisso può essere personalizzato in funzione delle proprie esigenze. Con questa opzione è necessario indicare il numero di campi da creare, indipendentemente dal numero di righe dei dati originali. Per esempio, se **Numero di campi nuovi** è impostato su 100, tutti i dati oltre le prime 100 righe saranno scartati. Se i dati originali contengono meno di 100 righe, alcuni campi conterranno valori null. Il numero dei campi può essere aumentato a piacere, ma lo scopo di questa impostazione è evitare la trasposizione di un milione di record in un milione di campi, che genererebbe un risultato impossibile da gestire.

Si supponga per esempio di disporre di dati con serie suddivise per righe e di un campo separato (colonna) per ogni mese. Questi dati possono essere trasposti in modo che ogni serie si trovi in un campo separato, con una riga per ogni mese.

Leggi dal campo. Legge i nomi dei campi da un campo esistente. Con questa opzione, il numero dei nuovi campi è determinato dai dati, fino al limite massimo specificato. Ogni valore del campo selezionato diventa un nuovo campo nei dati di output. Il campo selezionato può avere qualsiasi tipo di archiviazione (numero intero, stringa, data e così via) ma per evitare che due campi abbiano lo stesso nome, ogni valore del campo selezionato deve essere univoco (in altre parole, il numero dei valori deve corrispondere al numero delle righe). Se vengono rilevati campi con nomi uguali, viene visualizzato un avviso.

- **Leggi valori.** Se il campo selezionato non è stato istanziato, selezionare questa opzione per creare l'elenco dei nuovi nomi di campi. Se il campo è già stato istanziato, questa operazione non è necessaria.
- **Numero massimo di valori da leggere.** Quando i nomi dei campi vengono letti dai dati viene indicato un limite superiore per evitare di creare un numero eccessivo di campi (come notato in precedenza, la trasposizione di un milione di record in un milione di campi genererebbe un risultato impossibile da gestire).

Per esempio, se la prima colonna nei dati specifica il nome di ciascuna serie, è possibile usare questi valori come nomi di campi nei dati trasposti.

Trasponi. Per default, vengono trasposti solo i campi continui (intervalli numerici) (archiviazione di numeri interi o reali). Se lo si desidera, è possibile scegliere un sottoinsieme di campi numerici o trasporre invece solo campi stringa. Tuttavia, tutti i campi trasposti devono avere lo stesso tipo di archiviazione — numerica o stringa, ma non entrambi — poiché se i campi di input vengono mischiati, vengono generati valori misti all'interno di ciascuna colonna di output, in violazione della regola secondo cui tutti i valori di un campo devono avere lo stesso tipo di archiviazione. Gli altri tipi di archiviazione (data, ora, timestamp) non possono essere trasposti.

- **Tutti i numerici.** Traspone tutti i campi numerici (archiviazione di numeri interi o reali). Il numero delle righe di output corrisponde al numero di campi numerici presente nei dati originali.
- **Tutte le stringhe.** Traspone tutti i campi stringa.
- **Personalizzato.** Consente di selezionare un sottoinsieme di campi numerici. Il numero delle righe di output corrisponde al numero dei campi selezionati. *Nota:* questa opzione è disponibile solo per i campi numerici.

Nome ID riga. Specifica il nome del campo ID riga creato dal nodo. I valori di questo campo sono determinati dai nomi dei campi dei dati originali.

Suggerimento: quando si esegue la trasposizione dei dati di serie temporali da righe a colonne, se i dati originali includono una riga, per esempio data, mese o anno, che etichetta il periodo relativo a ogni misurazione, assicurarsi di leggere queste etichette in IBM SPSS Modeler come nomi di campi (come illustrato negli esempi precedenti, che mostrano il mese o la data come nomi di campi nei dati originali) invece di includere l'etichetta nella prima riga di dati. In questo modo si eviterà di mescolare etichette e valori nelle singole colonne (in quel caso, i numeri verrebbero letti come stringhe dato che non ci possono essere più tipi di archiviazione all'interno di una stessa colonna).

Nodo Intervalli di tempo

Il nodo Intervalli di tempo consente di specificare gli intervalli e generare le etichette per i dati di serie temporali da utilizzare in un nodo di modellazione Serie temporali, Streaming TS o Grafico temporale per la stima o la previsione. È supportata una gamma completa di intervalli temporali, dai secondi agli anni. Per esempio, se si dispone di una serie di misurazioni giornaliere che iniziano il 3 gennaio 2005, è possibile assegnare delle etichette ai record a partire da quella data; la seconda riga sarà quella del 4 gennaio, e così via. È anche possibile specificare la periodicità — ad esempio, cinque giorni alla settimana oppure otto ore al giorno.

Inoltre, è possibile specificare l'intervallo dei record da utilizzare per le stime. Si può scegliere di escludere i primi record della serie e di specificare i record di holdout. In questo modo sarà possibile testare il modello utilizzando come controllo i record più recenti delle serie temporali per confrontare i loro valori noti con i valori stimati per quei periodi.

È inoltre possibile specificare quanti periodi di tempo in futuro si desidera prevedere, ed i valori futuri utilizzati dai nodi downstream di modellazione Serie temporali o Streaming TS nella previsione.

Il nodo Intervalli di tempo genera un campo *EtichettaTempo* in un formato appropriato per l'intervallo e il periodo indicati, insieme a un campo *IndiceTempo* che assegna un numero intero univoco a ogni record. È possibile che vengano generati anche diversi campi aggiuntivi, a seconda del periodo o dell'intervallo selezionato (per esempio, il minuto o il secondo in cui avviene una misurazione).

È possibile riempire o aggregare i valori in funzione delle proprie esigenze per fare in modo che le misurazioni siano ripartite in modo uniforme. I metodi di modellazione dei dati di serie temporali richiedono un intervallo uniforme tra una misurazione e l'altra, con gli eventuali valori mancanti indicati da righe vuote. Se i dati a disposizione dell'utente non possiedono già queste caratteristiche, il nodo è in grado di trasformarli nel formato richiesto.

Commenti

- È possibile che gli intervalli periodici non corrispondano al tempo reale. Per esempio, una serie basata su una settimana lavorativa standard di cinque giorni considererebbe l'interruzione fra venerdì e lunedì come un solo giorno.
- Il nodo Intervalli di tempo presuppone che ogni serie si trovi in un campo o in una colonna, con una riga per ogni misurazione. Se necessario, è possibile trasporre i dati per soddisfare questo requisito. Per ulteriori informazioni, consultare l'argomento "Nodo Trasponi" a pagina 150.

- Per le serie non ripartite uniformemente, è possibile specificare un campo che identifica la data o l'ora di ogni misurazione. Si noti che sono richiesti un campo data, ora o timestamp nel formato appropriato da utilizzare come input. Se necessario, è possibile convertire un campo esistente (come un campo etichetta stringa) in questo formato utilizzando un nodo Riempimento. Per ulteriori informazioni, consultare l'argomento "Conversione della classe di archiviazione mediante il nodo Riempimento" a pagina 133.
- Quando vengono visualizzati i dettagli dei campi delle etichette e dell'indice che sono stati generati, in molti casi può essere utile attivare la visualizzazione delle etichette dei valori. Per esempio, quando si esamina una tabella contenente dei valori generati per dati mensili, è possibile fare clic sull'icona delle etichette dei valori nella barra degli strumenti per visualizzare *gennaio*, *febbraio*, *marzo* e così via anziché *1*, *2*, *3* e così via.



Figura 6. Icona etichette valori

Indicazione degli intervalli di tempo

La scheda Intervalli consente di indicare l'intervallo e la periodicità per la creazione della serie o l'assegnazione di etichette alla stessa. Le impostazioni specifiche dipendono dall'intervallo selezionato. Per esempio, se si sceglie **Ore al giorno**, è possibile indicare il numero di giorni alla settimana, il giorno d'inizio di ogni settimana, il numero di ore al giorno e l'ora d'inizio di ogni giorno. Per ulteriori informazioni, consultare l'argomento "Intervalli supportati" a pagina 155.

Assegnazione di etichette o creazione della serie

È possibile identificare i record consecutivamente o creare le serie in base a un campo data, timestamp o ora specifico.

- **Iniziare assegnando un'etichetta al primo record.** Indicare la data e/o l'ora d'inizio per assegnare l'etichetta ai record consecutivi. Per assegnare l'etichetta Ore al giorno, per esempio, è necessario specificare la data e l'ora d'inizio della serie, che deve proseguire con un solo record per ogni ora successiva. A parte l'aggiunta delle etichette, questo metodo non modifica i dati originali. Esso presuppone invece che i record siano già ripartiti in modo equivalente, con un intervallo uniforme tra tutte le misurazioni. Le eventuali misurazioni mancanti devono essere indicate nei dati da righe vuote.
- **Crea dai dati.** Per le serie non ripartite uniformemente, è possibile specificare un campo che identifica la data o l'ora di ogni misurazione. Si noti che sono richiesti un campo data, ora o timestamp nel formato appropriato da utilizzare come input. Per esempio, se si dispone di un campo stringa con valori come *Gen 2000*, *Feb 2000* e così via, è possibile convertirlo in campo data mediante un nodo Riempimento. Per ulteriori informazioni, consultare l'argomento "Conversione della classe di archiviazione mediante il nodo Riempimento" a pagina 133. L'opzione **Crea dai dati** trasforma i dati in modo che corrispondano all'intervallo specificato inserendo o aggregando i record in base alle necessità, ad esempio "facendo confluire" le settimane in mesi oppure sostituendo i record mancanti con valori vuoti o estrapolati. Le funzioni da utilizzare per riempire o aggregare i record possono essere impostate nella scheda Creazione. Per ulteriori informazioni, consultare l'argomento "Opzioni di creazione degli intervalli di tempo" a pagina 153.

Nuova estensione nome di campo. Consente di specificare un prefisso o un suffisso da applicare a tutti i campi generati dal nodo. Ad esempio, utilizzando il prefisso di default *\$TI_*, i campi creati dal nodo vengono denominati *\$TI_TimeIndex*, *\$TI_TimeLabel* e così via.

Formato data. Indica il formato del campo *EtichettaTempo* creato dal nodo applicabile all'intervallo corrente. La disponibilità di questa opzione dipende dalla selezione corrente.

Formato ora. Indica il formato del campo *EtichettaTempo* creato dal nodo applicabile all'intervallo corrente. La disponibilità di questa opzione dipende dalla selezione corrente.

Opzioni di creazione degli intervalli di tempo

La scheda Creazione del nodo Intervalli di tempo consente di selezionare le opzioni per l'aggregazione o il riempimento dei campi al fine di farli corrispondere all'intervallo specificato. Queste impostazioni sono valide solo quando nella scheda Intervalli è selezionata l'opzione **Crea dai dati**. Per esempio, se si dispone di un insieme eterogeneo di dati settimanali e mensili, è possibile aggregare o "accorpare" i valori settimanali per ottenere un intervallo mensile uniforme. In alternativa è possibile impostare un intervallo settimanale e riempire la serie inserendo valori vuoti per tutte le settimane mancanti o estrapolando i valori mancanti mediante una funzione di padding specifica.

In caso di riempimento o aggregazione dei dati, tutti i campi data o timestamp esistenti vengono sostituiti a tutti gli effetti dai campi *EtichettaTempo* e *IndiceTempo* generati e vengono esclusi dall'output. Anche i campi senza tipo vengono eliminati. I campi che misurano il tempo come durata vengono conservati — ad esempio, i campi che misurano la lunghezza di una chiamata di servizio piuttosto che l'ora di avvio della chiamata — se sono archiviati interamente come campi ora piuttosto che come timestamp. Per ulteriori informazioni, consultare l'argomento "Impostazione dell'archiviazione e della formattazione dei campi" a pagina 24. Gli altri campi vengono aggregati in base alle opzioni specificate nella scheda Creazione.

- **Utilizza campi e funzioni di default.** Indica che tutti i campi devono essere opportunamente aggregati o riempiti, fatta eccezione per i campi data, timestamp e per quelli senza tipo come notato in precedenza. La funzione di default viene applicata in base al livello di misurazione — ad esempio, i campi continui vengono aggregati utilizzando la media, mentre i campi nominali utilizzano la modalità. L'impostazione di default di uno o più livelli di misurazione può essere modificata nella parte inferiore della finestra di dialogo.
- **Specifica campi e funzioni.** Consente di specificare i campi da riempire o da aggregare e la funzione da utilizzare per ciascuna operazione. I campi non selezionati vengono eliminati dall'output. Aggiungere o eliminare i campi dalla tabella mediante le icone a destra, oppure fare clic nella cella all'interno della colonna desiderata per modificare la funzione di aggregazione o di padding utilizzata per un determinato campo in modo da ignorare l'impostazione di default. I campi senza tipo sono esclusi dall'elenco e non possono essere aggiunti alla tabella.

Default. Indica le funzioni di aggregazione e di padding utilizzate per default per i vari tipi di campi. Queste impostazioni di default vengono applicate quando è selezionato **Utilizza impostazioni di default** e anche come impostazioni di default iniziali per tutti i nuovi campi aggiunti alla tabella. La modifica delle impostazioni di default non modifica le impostazioni esistenti della tabella ma è valida per tutti i campi aggiunti in seguito.

Funzione di aggregazione. Sono disponibili le seguenti funzioni di aggregazione:

- **Continuo.** Le funzioni disponibili per i campi continui sono **Media**, **Somma**, **Moda**, **Min** e **Max**.
- **Nominale.** Le opzioni disponibili sono **Moda**, **Primi** e **Ultimo**. **Primi** indica il primo valore non null (ordinato per data) del gruppo di aggregazione; **Ultimo** indica l'ultimo valore non null del gruppo.
- **Flag.** Le opzioni disponibili sono **Vero se uno qualsiasi è vero**, **Moda**, **Primi** e **Ultimo**.

Funzione di padding. Sono disponibili le seguenti funzioni di padding:

- **Continuo.** Le opzioni disponibili sono **Vuoto** e **Media dei punti recenti**, ovvero la media dei tre valori non null più recenti precedenti al periodo di tempo che sarà creato. Se non sono disponibili tre valori, il nuovo valore è vuoto. I valori recenti comprendono solo i valori effettivi; i valori riempiti creati in precedenza non vengono considerati nella ricerca di un valore non null.
- **Nominale. Vuoto e Valore più recente.** "Più recente" si riferisce al valore non null più recente precedente al periodo di tempo che sarà creato. Anche in questo caso, nella ricerca di un valore recente vengono considerati solo i valori effettivi.

- **Flag.** Le opzioni disponibili sono **Vuoto**, **Vero** e **Falso**.

Numero massimo di record in un insieme di dati risultante. Consente di impostare un limite superiore al numero di record creati che potrebbe altrimenti aumentare a dismisura, soprattutto quando l'intervallo di tempo è impostato (deliberatamente o meno) sui secondi. Per esempio, una serie comprendente solo due valori (1 gennaio 2000 e 1 gennaio 2001) genererebbe 31.536.000 record se venisse riempita in base ai secondi (60 secondi x 60 minuti x 24 ore x 365 giorni). Quando si supera il limite massimo impostato, il sistema interrompe l'elaborazione e visualizza un avviso.

Campo Conteggio

Quando i valori vengono aggregati o riempiti, il sistema crea un nuovo campo denominato *Conteggio* che indica il numero di record utilizzati per determinare il nuovo record. Per esempio, se in un solo mese vengono aggregati quattro valori settimanali il conteggio sarebbe pari a 4. Per un record riempito, il conteggio è 0. Il nome del campo è *Conteggio* più il prefisso o il suffisso specificato nella scheda Intervallo.

Periodo di stima

La scheda Stima del nodo Intervalli di tempo consente di specificare la gamma di record utilizzati nella stima del modello, oltre agli eventuali record di holdout. Queste impostazioni possono essere sovrascritte nei nodi Modelli a valle, ma la specifica delle stesse a questo punto potrebbe essere più comoda che la specifica a livello dei singoli nodi.

Inizia stima. È possibile far partire il periodo di stima all'inizio dei dati o escludere i valori meno recenti che potrebbero risultare poco utili ai fini della previsione. A seconda dei dati, la riduzione del periodo di stima potrebbe migliorare la performance e ridurre il tempo necessario alla preparazione dei dati, senza comportare una significativa diminuzione della precisione della previsione.

Termina stima. È possibile eseguire una stima del modello utilizzando tutti i record fino alla fine dei dati oppure utilizzare i record più recenti come record di controllo per valutare il modello. Nel secondo caso, si sta eseguendo la "previsione" di valori già noti, il che consente di confrontare i valori osservati e predittivi per misurare l'efficacia del modello.

Previsioni

La scheda Previsione del nodo Intervalli di tempo consente di specificare il numero di record che si desidera prevedere e di specificare i valori futuri utilizzati nelle previsioni dai nodi downstream di modellazione Serie temporali o Streaming TS. Queste impostazioni possono essere sovrascritte nei nodi downstream in base alle esigenze ma specificarli qui può essere più comodo che specificarli singolarmente per ogni nodo.

Estendi record nel futuro. Specifica il numero dei record da prevedere oltre il periodo di stima. Si noti che questi record possono essere "previsti" o meno, a seconda del numero dei record di holdout specificati nella scheda Stima.

Campo indicatore futuro. Etichetta del campo generato che indica se un record contiene i dati di previsione o meno. Il valore di default per l'etichetta è *\$TI_Futuro*.

Valori futuri da utilizzare nelle previsioni. Per ciascun record che si desidera prevedere (ad eccezione dei record di holdout), se si utilizzano campi predittori (con il ruolo impostato su *Input*), è necessario specificare i valori stimati per il periodo di previsione per ogni predittore. Questi valori possono essere specificati manualmente oppure scelti da un elenco.

- **Campo.** Fare clic sul pulsante del selettore di campo e scegliere gli eventuali campi che potrebbero essere utilizzati come predittori. Si noti che i campi selezionati qui possono essere utilizzati nella modellazione oppure no; per utilizzare effettivamente un campo come predittore, è necessario selezionarlo in un nodo Modelli a valle. Questa finestra di dialogo rappresenta semplicemente un

metodo conveniente per specificare i valori futuri in modo che possano essere condivisi da più nodi Modelli a valle senza che sia necessario specificarli separatamente nei singoli nodi. Si noti anche che l'elenco dei campi disponibili può essere limitato a seconda delle selezioni effettuate nella scheda Creazione. Per esempio, se nella scheda Creazione è stata selezionata l'opzione **Specifica campi e funzioni**, i campi non aggregati o riempiti vengono esclusi dal flusso e non possono essere utilizzati nella modellazione.

Nota: se vengono specificati valori futuri per un campo che non è più disponibile nel flusso (perché è stato escluso oppure a causa di nuove selezioni eseguite nella scheda Creazione), tale campo viene visualizzato in rosso nella scheda Previsione.

- **Valori.** Per ciascun campo, è possibile scegliere da un elenco di funzioni o fare clic su **Specifica** per immettere i valori manualmente o selezionare i valori predefiniti dall'elenco. Se i campi predittori sono correlati a elementi sotto il controllo dell'utente o comunque conoscibili in anticipo, immettere i valori manualmente. Per esempio, se si sta eseguendo la previsione delle entrate di un hotel per il mese prossimo sulla base del numero delle camere prenotate, è possibile specificare il numero delle prenotazioni effettive di cui già si dispone per quel periodo. Al contrario, se un campo predittore si riferisce a un elemento al di fuori del controllo dell'utente, per esempio al prezzo di un titolo azionario, è possibile utilizzare una funzione, quale il valore più recente oppure la media dei punti recenti. Le funzioni disponibili dipendono dal livello di misurazione del campo.

Tabella 26. Funzioni disponibili per i livelli di misurazione.

Livello di misurazione	Funzioni
Campo continuo o nominale	Blank Media dei punti recenti Valore più recente Specify
Campo flag	Blank Valore più recente True False Specify

Media dei punti recenti—Calcola il valore futuro in base alla media degli ultimi tre punti di dati.

Valore più recente—Imposta il valore futuro su quello del punto di dati più recente.

True/Falsw—Imposta il valore futuro di un campo flag su Vero o Falso, come specificato.

Specify—Visualizza una finestra di dialogo che consente di specificare i valori futuri manualmente o di selezionarli da un elenco predefinito.

Valori futuri

Qui è possibile specificare i valori futuri da utilizzare nelle previsioni con i nodi di modellazione downstream Serie temporali o Streaming TS. Queste impostazioni possono essere sovrascritte nei nodi downstream in base alle esigenze, ma specificarli qui può essere più comodo che specificarli singolarmente per ogni nodo.

È possibile immettere i valori manualmente, oppure fare clic sul pulsante di selezione a destra della finestra di dialogo per scegliere da un elenco di valori definiti per il campo corrente.

Il numero dei valori futuri che possono essere specificati corrisponde al numero dei record in base ai quali si sta estendendo la serie temporale nel futuro.

Intervalli supportati

Il nodo Intervalli di tempo supporta una gamma di intervalli completa, dai secondi agli anni, nonché periodi ciclici (per esempio, le stagioni) e non ciclici. L'intervallo viene specificato nel campo Intervallo di tempo nella scheda Intervalli.

Periodi

Selezionare l'opzione **Periodi** per assegnare un'etichetta a una serie esistente non ciclica che non corrisponde ad alcuno degli altri intervalli specificati. La serie deve già essere nell'ordine esatto, con un intervallo uniforme tra una misurazione e l'altra. Quando è selezionato questo intervallo, l'opzione **Crea dai dati** non è disponibile.

Output campione

L'assegnazione delle etichette ai record avviene in modo incrementale, in base al valore iniziale specifico (*Periodo 1, Periodo 2* e così via). Vengono creati nuovi campi come mostrato nella tabella riportata di seguito.

Tabella 27. Esempi di creazione di nuovi campi

\$TI_IndiceTempo (Intero)	\$TI_EtichettaTempo (Stringa)	\$TI_Periodo (Intero)
1	Periodo 1	1
2	Periodo 2	2
3	Periodo 3	3
4	Periodo 4	4
5	Periodo 5	5

Periodi ciclici

Selezionare **Periodi ciclici** per assegnare un'etichetta a una serie esistente caratterizzata da un ciclo ripetitivo che non corrisponde ad alcuno degli intervalli standard. Questa opzione si potrebbe utilizzare, per esempio, se un anno fiscale comprende solo 10 mesi. La serie deve già essere nell'ordine esatto, con un intervallo uniforme tra una misurazione e l'altra. Quando è selezionato questo intervallo, l'opzione **Crea dai dati** non è disponibile.

Output campione

L'assegnazione delle etichette ai record avviene in modo incrementale, in base al ciclo iniziale e al periodo indicato (*Ciclo 1, Periodo 1, Ciclo 1, Periodo 2* e così via). Ad esempio, con il numero di periodi per ciclo impostato su 3, i nuovi campi vengono creati come mostrato nella tabella riportata di seguito.

Tabella 28. Esempi di creazione di nuovi campi.

\$TI_IndiceTempo (Intero)	\$TI_EtichettaTempo (Stringa)	\$TI_Ciclo (Intero)	\$TI_Periodo (Intero)
1	Ciclo 1, Periodo 1	1	1
2	Ciclo 1, Periodo 2	1	2
3	Ciclo 1, Periodo 3	1	3
4	Ciclo 2, Periodo 1	2	1
5	Ciclo 2, Periodo 2	2	2

Anni

Per gli anni è possibile indicare l'anno di inizio per identificare una serie di record consecutivi, oppure selezionare **Crea dai dati** per specificare un campo timestamp o data che identifichi l'anno di ogni singolo record.

Output campione

Vengono creati nuovi campi come mostrato nella tabella riportata di seguito.

Tabella 29. Esempi di creazione di nuovi campi

\$TI-IndiceTempo (Intero)	\$TI-EtiquetaTempo (Stringa)	\$TI-Anno (Intero)
1	2000	2000
2	2001	2001
3	2002	2002
4	2003	2003
5	2004	2004

Trimestri

Per le serie trimestrali è possibile specificare il mese in cui inizia l'anno fiscale. È possibile inoltre indicare il trimestre e l'anno di inizio (per esempio, T1 2000) per identificare una serie di record consecutivi, oppure selezionare **Crea dai dati** per scegliere un campo timestamp o data che identifichi il trimestre e l'anno di ogni singolo record.

Output campione

Per un anno fiscale che inizia in gennaio, i nuovi campi vengono creati e popolati come illustrato nella tabella riportata di seguito.

Tabella 30. Esempi di creazione di nuovi campi

\$TI-IndiceTempo (Intero)	\$TI-EtiquetaTempo (Stringa)	\$TI-Anno (Intero)	\$TI-Trimestre (Intero con etichette)
1	T1 2000	2000	1 (T1)
2	T2 2000	2000	2 (T2)
3	T3 2000	2000	3 (T3)
4	T4 2000	2000	4 (T4)
5	Q1 2001	2001	1 (T1)

Se l'anno inizia in un mese diverso da gennaio, i nuovi campi sono come mostrato nella tabella riportata di seguito (supponendo un anno fiscale che inizia in luglio). Per visualizzare le etichette che identificano i mesi di ogni trimestre, attivare la visualizzazione delle etichette dei valori facendo clic sull'icona della barra degli strumenti.



Figura 7. Icona etichette valori

Tabella 31. Esempi di creazione di nuovi campi

\$TI-IndiceTempo (Intero)	\$TI-EtiquetaTempo (Stringa)	\$TI-Anno (Intero)	\$TI-Trimestre (Intero con etichette)
1	T1 2000/2001	1	1 (T1 Lug-Set)
2	T2 2000/2001	1	2 (T2 Ott-Dic)
3	T3 2000/2001	1	3 (T3 Gen-Mar)
4	T4 2000/2001	1	4 (T4 Apr-Giu)
5	Q1 2001/2002	2	1 (T1 Lug-Set)

Mesi

È possibile selezionare l'anno e il mese d'inizio per identificare una serie consecutiva di record oppure selezionare **Crea dai dati** per scegliere un campo timestamp o data che indichi il mese di ogni singolo record.

Output campione

Vengono creati nuovi campi come mostrato nella tabella riportata di seguito.

Tabella 32. Esempi di creazione di nuovi campi

\$TI-IndiceTempo (Intero)	\$TI-EtichettaTempo (Data)	\$TI-Anno (Intero)	\$TI-Mesi (Intero con etichette)
1	Gen 2000	2000	1 (Gennaio)
2	Feb 2000	2000	2 (Febbraio)
3	Mar 2000	2000	3 (Marzo)
4	Apr 2000	2000	4 (Aprile)
5	Mag 2000	2000	5 (Maggio)

Settimane (non periodiche)

Per una serie settimanale, è possibile selezionare il giorno della settimana in cui inizia il ciclo.

Si noti che le settimane possono essere soltanto non periodiche perché i diversi mesi, trimestri e anche gli anni non hanno necessariamente lo stesso numero di settimane. Tuttavia, i dati con timestamp possono essere facilmente aggregati o riempiti a un livello settimanale per i modelli non periodici.

Output campione

Vengono creati nuovi campi come mostrato nella tabella riportata di seguito.

Tabella 33. Esempi di creazione di nuovi campi

\$TI-IndiceTempo (Intero)	\$TI-EtichettaTempo (Data)	\$TI-Settimana (Intero)
1	1999-12-27	1
2	2000-01-03	2
3	2000-01-10	3
4	2000-01-17	4
5	2000-01-24	5

Il campo *\$TI-EtichettaTempo* per una settimana visualizza il primo giorno di quella settimana. Nella tabella precedente, l'utente inizia l'etichettatura dal primo gennaio del 2000. Tuttavia, la settimana inizia di lunedì mentre il primo gennaio 2000 è un sabato. Pertanto, la settimana che comprende il primo gennaio inizia il 27 dicembre 1999 e rappresenta l'etichetta del primo punto.

Il formato data determina le stringhe prodotte per il campo *\$TI-EtichettaTempo*.

Giorni a settimana

Per le misurazioni quotidiane che avvengono in un ciclo settimanale, è possibile specificare il numero di giorni a settimana e il giorno d'inizio di ogni settimana. È possibile indicare una data d'inizio per identificare una serie consecutiva di record oppure selezionare **Crea dai dati** per scegliere un campo timestamp o data che indichi la data di ogni singolo record.

Output campione

Vengono creati nuovi campi come mostrato nella tabella riportata di seguito.

Tabella 34. Esempi di creazione di nuovi campi

\$TI-IndiceTempo (Intero)	\$TI-EtichettaTempo (Data)	\$TI-Settimana (Intero)	\$TI-Giorno (Intero con etichette)
1	5 Gen 2005	1	3 (Mercoledì)
2	6 Gen 2005	1	4 (Giovedì)
3	7 Gen 2005	1	5 (Venerdì)
4	10 Gen 2005	2	1 (Lunedì)
5	11 Gen 2005	2	2 (Martedì)

Nota: la settimana inizia sempre da 1 per il primo periodo di tempo e non segue il ciclo del calendario. Pertanto, la 52° settimana è seguita dalla 53°, dalla 54° e così via. La settimana non corrisponde alla settimana dell'anno, ma solo al numero di incrementi settimanali della serie.

Giorni (non periodici)

Scegliere i giorni non periodici in presenza di misurazioni quotidiane che non rientrano in un normale ciclo settimanale. È possibile indicare una data d'inizio per identificare una serie consecutiva di record oppure selezionare **Crea dai dati** per scegliere un campo timestamp o data che indichi la data di ogni singolo record.

Output campione

Vengono creati nuovi campi come mostrato nella tabella riportata di seguito.

Tabella 35. Esempi di creazione di nuovi campi

\$TI-IndiceTempo (Intero)	\$TI-EtichettaTempo (Data)
1	5 Gen 2005
2	6 Gen 2005
3	7 Gen 2005
4	8 Gen 2005
5	9 Gen 2005

Ore al giorno

Per le misurazioni orarie che rientrano in un ciclo giornaliero è possibile specificare il numero di giorni a settimana, il numero di ore al giorno (per esempio una giornata lavorativa di otto ore), il giorno d'inizio della settimana e l'ora d'inizio di ogni giorno. Le ore possono essere indicate con precisione al minuto in base al formato 24 ore (per esempio, 14:05 = 2:05 p.m.)

È possibile indicare la data e l'ora d'inizio per identificare una serie consecutiva di record oppure selezionare **Crea dai dati** per scegliere un campo timestamp che identifichi la data e l'ora di ogni singolo record.

Output campione

Vengono creati nuovi campi come mostrato nella tabella riportata di seguito.

Tabella 36. Esempi di creazione di nuovi campi

\$TI-IndiceTempo (Intero)	\$TI-EtichettaTempo (Timestamp)	\$TI-Giorno (Intero con etichette)	\$TI-Ora (Intero con etichette)
1	5 Gen 2005 8:00	3 (Mercoledì)	8 (8:00)

Tabella 36. Esempi di creazione di nuovi campi (Continua)

\$TI-IndiceTempo (Intero)	\$TI-EtichettaTempo (Timestamp)	\$TI-Giorno (Intero con etichette)	\$TI-Ora (Intero con etichette)
2	5 Gen 2005 9:00	3 (Mercoledì)	9 (9:00)
3	5 Gen 2005 10:00	3 (Mercoledì)	10 (10:00)
4	5 Gen 2005 11:00	3 (Mercoledì)	11 (11:00)
5	5 Gen 2005 12:00	3 (Mercoledì)	12 (12:00)

Ore (non periodiche)

Scegliere questa opzione in presenza di misurazioni orarie che non rientrano in un normale ciclo quotidiano. È possibile indicare l'ora d'inizio per identificare una serie consecutiva di record oppure selezionare **Crea dai dati** per scegliere un campo timestamp o ora che indichi l'ora di ogni singolo record.

Le ore si basano sul formato 24 ore (13:00 = 1:00 p.m.) e non ripartono dall'inizio (l'ora 24 è seguita dall'ora 25).

Output campione

Vengono creati nuovi campi come mostrato nella tabella riportata di seguito.

Tabella 37. Esempi di creazione di nuovi campi

\$TI-IndiceTempo (Intero)	\$TI-EtichettaTempo (Stringa)	\$TI-Ora (Intero con etichette)
1	8:00	8 (8:00)
2	9:00	9 (9:00)
3	10:00	10 (10:00)
4	11:00	11 (11:00)
5	12:00	12 (12:00)

Minuti al giorno

Per le misurazioni al minuto che rientrano in un ciclo quotidiano, è possibile specificare il numero di giorni alla settimana, il giorno d'inizio della settimana, il numero di ore al giorno e l'ora d'inizio del giorno. Le ore vengono indicate in base al formato 24 ore e si possono specificare con una precisione di minuti e secondi utilizzando i due punti (per esempio, 2:05:17 p.m. = 14:05:17). È possibile anche specificare il numero di minuti per gli incrementi (ogni minuto, ogni due minuti e così via; l'incremento deve essere un valore divisibile per 60).

È possibile indicare la data e l'ora d'inizio per identificare una serie consecutiva di record oppure selezionare **Crea dai dati** per scegliere un campo timestamp che identifichi la data e l'ora di ogni singolo record.

Output campione

Vengono creati nuovi campi come mostrato nella tabella riportata di seguito.

Tabella 38. Esempi di creazione di nuovi campi

\$TI-IndiceTempo (Intero)	\$TI-EtichettaTempo (Timestamp)	\$TI-Minuto
1	2005-01-05 08:00:00	0
2	2005-01-05 08:01:00	1
3	2005-01-05 08:02:00	2

Tabella 38. Esempi di creazione di nuovi campi (Continua)

\$TI-IndiceTempo (Intero)	\$TI-EtichettaTempo (Timestamp)	\$TI-Minuto
4	2005-01-05 08:03:00	3
5	2005-01-05 08:04:00	4

Minuti (non periodici)

Scegliere questa opzione in presenza di misurazioni al minuto che non rientrano in un normale ciclo quotidiano. È possibile specificare il numero di minuti per gli incrementi (ogni minuto, ogni due minuti e così via; il valore indicato deve essere un numero divisibile per 60).

È possibile indicare l'ora d'inizio per identificare una serie consecutiva di record oppure selezionare **Crea dai dati** per scegliere un campo timestamp o ora che identifichi l'ora di ogni singolo record.

Output campione

Vengono creati nuovi campi come mostrato nella tabella riportata di seguito.

Tabella 39. Esempi di creazione di nuovi campi

\$TI-IndiceTempo (Intero)	\$TI-EtichettaTempo (Stringa)	\$TI-Minuto
1	8:00	0
2	8:01	1
3	8:02	2
4	8:03	3
5	8:04	4

- La stringa *EtichettaTempo* viene creata utilizzando i due punti per separare le ore dai minuti. L'ora non riparte dall'inizio — l'ora 24 è seguita dall'ora 25.
- I minuti vengono incrementati in base al valore impostato nella finestra di dialogo. Per esempio, se l'incremento è 2, il valore di *EtichettaTempo* sarà 8:00, 8:02 e così via, e i minuti saranno 0, 2 e così via.

Secondi al giorno

Per gli intervalli di secondi che rientrano in un ciclo quotidiano, è possibile specificare il numero di giorni alla settimana, il giorno d'inizio della settimana, il numero di ore al giorno e l'ora d'inizio del giorno. Le ore vengono indicate in base al formato 24 ore e si possono specificare con una precisione di minuti e secondi utilizzando i due punti (per esempio, 2:05:17 p.m. = 14:05:17). È possibile anche specificare il numero di secondi per gli incrementi (ogni secondo, ogni due secondi e così via; il valore indicato deve essere un numero divisibile per 60).

È possibile specificare la data e l'ora d'inizio per identificare una serie consecutiva di record oppure selezionare **Crea dai dati** per scegliere un campo timestamp che indichi la data e l'ora di ogni singolo record.

Output campione

Vengono creati nuovi campi come mostrato nella tabella riportata di seguito.

Tabella 40. Esempi di creazione di nuovi campi

\$TI-IndiceTempo (Intero)	\$TI-EtichettaTempo (Timestamp)	\$TI-Minuto	\$TI-Secondo
1	2005-01-05 08:00:00	0	0
2	2005-01-05 08:00:01	0	1
3	2005-01-05 08:00:02	0	2

Tabella 40. Esempi di creazione di nuovi campi (Continua)

\$TI-IndiceTempo (Intero)	\$TI-EtichettaTempo (Timestamp)	\$TI-Minuto	\$TI-Secondo
4	2005-01-05 08:00:03	0	3
5	2005-01-05 08:00:04	0	4

Secondi (non periodici)

Scegliere questa opzione in presenza di misurazioni eseguite al secondo che non rientrano in un normale ciclo quotidiano. È possibile specificare il numero di secondi per gli incrementi (ogni secondo, ogni due secondi e così via; il valore indicato deve essere un numero divisibile per 60).

Specificare l'ora d'inizio per identificare una serie consecutiva di record oppure selezionare **Crea dai dati** per scegliere un campo timestamp o ora che identifichi l'ora di ogni singolo record.

Output campione

Vengono creati nuovi campi come mostrato nella tabella riportata di seguito.

Tabella 41. Esempi di creazione di nuovi campi

\$TI-IndiceTempo (Intero)	\$TI-EtichettaTempo (Stringa)	\$TI-Minuto	\$TI-Secondo
1	8:00:00	0	0
2	8:00:01	0	1
3	8:00:02	0	2
4	8:00:03	0	3
5	8:00:04	0	4

- La stringa *EtichettaTempo* viene creata utilizzando i due punti per separare le ore dai minuti e i minuti dai secondi. L'ora non riparte da zero — l'ora 24 è seguita dall'ora 25.
- I secondi vengono incrementati del numero indicato come incremento. Se l'incremento è 2, il valore di *EtichettaTempo* sarà 08:00:00, 08:00:02 e così via, e i secondi saranno 0, 2 e così via.

Nodo Cronologia

I nodi Cronologia sono utilizzati in genere per dati sequenziali, per esempio per dati di serie temporali. Consentono di creare nuovi campi contenenti dati dei campi di record precedenti. Quando si utilizza un nodo Cronologia, è possibile che si desideri preordinare i dati in base a un determinato campo. A questo scopo, è possibile utilizzare un nodo Ordina.

Impostazione delle opzioni relative al nodo Cronologia

Campi selezionati. Il Selettore di campo (campo a destra della casella di testo) permette di selezionare i campi per i quali si desidera esaminare la cronologia. Ogni campo selezionato viene utilizzato per creare nuovi campi per tutti i record nell'insieme di dati.

Offset. Specificare il record che precede quello corrente dal quale si desidera estrarre i valori di campo cronologici. Se Offset è impostato su 3, per esempio, al passaggio di ciascun record in questo nodo verranno inclusi nel record corrente i valori di campo del terzo record precedente. Utilizzare le impostazioni Estensione per specificare il punto a partire dal quale verranno estratti i record. Utilizzare le frecce per modificare il valore dell'offset.

Estensione. Specificare il numero di record precedenti a partire dal quale si desidera estrarre i valori. Per esempio, se Offset è impostato su 3 ed Estensione su 5, a ciascun record che passa per il nodo verranno

aggiunti cinque campi per ciascun campo specificato nell'elenco dei campi selezionati. Ciò significa che quando il nodo elabora il record 10, verranno aggiunti i campi dal record 7 al record 3. Utilizzare le frecce per modificare il valore dell'estensione.

Se la cronologia non è disponibile. Selezionare una delle opzioni seguenti per gestire i record senza valori cronologici. In genere si tratta dei record presenti all'inizio dell'insieme di dati, per i quali non esistono record precedenti da utilizzare come cronologia.

- **Scarta record.** Selezionare questa opzione per scartare i record nel caso in cui per il campo selezionato non esista un valore cronologico.
- **Lascia cronologia non definita.** Selezionare questa opzione per mantenere i record in cui non sia disponibile un valore cronologico. Il campo cronologico verrà riempito con un valore non definito, visualizzato come `$null$`.
- **Riempi valori con.** Specificare il valore o la stringa da utilizzare per i record in cui non sia disponibile un valore cronologico. Il valore sostitutivo di default è `undef`, il valore null di sistema. I valori null sono visualizzati utilizzando la stringa `$null$`.

Quando si seleziona un valore sostitutivo, per garantire un'esecuzione corretta è necessario tenere presente le seguenti regole:

- I campi selezionati devono essere dello stesso tipo di archiviazione.
- Se tutti i campi selezionati includono un'archiviazione numerica, il valore sostitutivo deve essere analizzato come numero intero.
- Se tutti i campi selezionati includono un'archiviazione reale, il valore sostitutivo deve essere analizzato come numero reale.
- Se tutti i campi selezionati includono un'archiviazione simbolica, il valore sostitutivo deve essere analizzato come stringa.
- Se tutti i campi selezionati includono un'archiviazione di data/ora, il valore sostitutivo deve essere analizzato come campo data/ora.

Se qualcuna delle condizioni sopra riportate non viene soddisfatta, durante l'esecuzione del nodo Cronologia verrà generato un errore.

Nodo Ordina campi

Il nodo Ordina campi consente di definire l'ordine naturale utilizzato per visualizzare i campi a valle. Tale ordine incide sulla visualizzazione dei campi in vari contesti, quali tabelle, elenchi e Selettore di campo. Questa operazione risulta utile, per esempio, se si desidera evidenziare i campi più interessanti negli insiemi di dati di grandi dimensioni.

Impostazione delle opzioni Ordina campi

I campi possono essere riordinati in due modi: ordinamento personalizzato e ordinamento automatico.

Ordinamento personalizzato

Selezionare **Ordinamento personalizzato** per attivare una tabella con i nomi e i tipi di campo in cui è possibile visualizzare tutti i campi e utilizzare i pulsanti con le frecce per creare un ordine personalizzato.

Per riordinare i campi:

1. Selezionare un campo nella tabella. Utilizzare il metodo Ctrl-clic per selezionare più campi.
2. Utilizzare i pulsanti con le frecce per spostare i campi verso l'alto o verso il basso di una riga.
3. Utilizzare i tasti direzionali per spostare i campi all'inizio o alla fine dell'elenco.
4. Specificare l'ordine dei campi non inclusi in questo contesto spostando la riga di divisione verso l'alto o verso il basso, indicata come [altri campi].

Ulteriori informazioni su [altri campi]

Altri campi. La riga di divisione [altri campi] consente di suddividere la tabella in due parti.

- I campi visualizzati al di sopra della riga di divisione verranno posizionati (in base alla visualizzazione nella tabella) prima di tutti gli ordini naturali utilizzati per visualizzare i campi a valle di questo nodo.
- I campi visualizzati al di sotto della riga di divisione verranno posizionati (in base alla visualizzazione nella tabella) dopo tutti gli ordini naturali utilizzati per visualizzare i campi a valle di questo nodo.

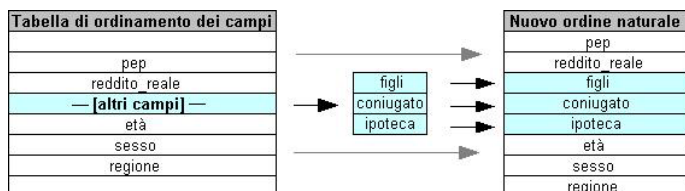


Figura 8. Diagramma che illustra il modo in cui gli "altri campi" sono incorporati nel nuovo ordine dei campi

- Tutti gli altri campi non visualizzati nella tabella di riordinamento dei campi verranno visualizzati tra i campi "superiore" ed "inferiore" come indicato dal posizionamento della riga di divisione.

Sono disponibili altre opzioni di ordinamento personalizzate, tra cui:

- Ordinare i campi in modo crescente o decrescente facendo clic sulle frecce poste sopra ogni intestazione di colonna (**Tipo**, **Nome** e **Classe archiviazione**). Quando si ordina in base alle colonne, i campi non specificati in questo punto (indicati dalla riga [altri campi]) vengono ordinati per ultimi in base all'ordine naturale.
- Fare clic su **Cancella non utilizzati** per eliminare tutti i campi non utilizzati dal nodo Ordina campi. I campi non utilizzati vengono visualizzati nella tabella con un carattere rosso. Ciò indica che il campo è stato eliminato in operazioni upstream.
- Specificare l'ordinamento per tutti i campi nuovi (contrassegnati da una saetta per indicare un campo nuovo o non specificato). Se si fa clic su **OK** o **Applica**, l'icona scompare.

Nota: se vengono aggiunti campi a monte dopo l'applicazione di un ordine personalizzato, i nuovi campi verranno accodati alla fine dell'elenco personalizzato.

Ordinamento automatico

Selezionare **Ordinamento automatico** per specificare un parametro per l'ordinamento. Le opzioni della finestra di dialogo cambiano in modo dinamico per mostrare opzioni relative all'ordinamento automatico.

Ordina per. Selezionare uno dei tre metodi di ordinamento dei campi disponibili nel nodo Riordina. I pulsanti con le frecce indicano se l'ordine sarà crescente o decrescente. Selezionare uno dei modi per eseguire una modifica.

- Nome
- Tipo
- Archiviazione

I campi aggiunti a monte del nodo Ordina campi dopo l'applicazione dell'ordinamento automatico verranno disposti automaticamente nella posizione corretta in base al tipo di ordinamento selezionato.

Capitolo 5. Nodi Grafico

Caratteristiche comuni dei nodi Grafico

Numerose fasi del processo di data mining utilizzano grafici e diagrammi per esaminare i dati immessi in IBM SPSS Modeler. Per esempio, è possibile connettere un nodo grafico o distribuzione a una sorgente dati per ottenere informazioni su distribuzioni e tipi di dati ed eseguire quindi manipolazioni su record e campi per preparare i dati per operazioni di creazione di modelli a valle. I grafici vengono comunemente utilizzati anche per controllare la distribuzione e le relazioni tra i nuovi campi derivati.

La palette Grafici contiene i seguenti nodi:



Il nodo Lavagna grafica offre numerosi tipi di grafici diversi in un unico nodo. Con questo nodo è possibile scegliere i campi di dati da esplorare e selezionare quindi un grafico fra quelli disponibili per i dati selezionati. Il nodo esclude automaticamente tutti i tipi di grafici non adatti ai campi selezionati.



Il nodo Plot mostra la relazione tra campi numerici. È possibile creare un grafico utilizzando punti (un grafico a dispersione) oppure linee.



Il nodo distribuzione mostra l'occorrenza di valori simbolici (categoriali), per esempio tipo o genere di ipoteca. In genere è possibile utilizzare un nodo distribuzione per mostrare squilibri nei dati, che possono essere successivamente corretti con un nodo bilanciamento prima di creare un modello.



Il nodo Istogramma mostra l'occorrenza dei valori per i campi numerici. Viene spesso utilizzato per analizzare i dati prima delle manipolazioni e della creazione del modello. Come il nodo distribuzione, anche il nodo Istogramma viene frequentemente utilizzato per rivelare squilibri nei dati.



Il nodo Raccolta mostra la distribuzione dei valori di un campo numerico in relazione ai valori di un altro, ovvero crea grafici simili a istogrammi. È utile per illustrare una variabile o un campo i cui valori vengono modificati nel tempo. La grafica 3-D consente inoltre di includere un asse simbolico che visualizza le distribuzioni per categoria.



Un nodo Multiplot crea un grafico che consente di visualizzare più campi Y in un singolo campo X. I campi Y sono rappresentati come linee colorate e ognuno di essi equivale a un nodo Plot con lo Stile impostato su **Linea** e la Modalità X impostata su **Ordina**. I multiplot sono utili quando si desidera esplorare la fluttuazione di numerose variabili nel tempo.



Il nodo Web illustra l'intensità della relazione tra valori di due o più campi simbolici (categoriali). Il grafico utilizza linee di spessore diverso per indicare l'intensità della connessione. Un nodo Web può essere utilizzato, per esempio, per analizzare la relazione tra l'acquisto di vari oggetti in un sito di e-commerce.



Il nodo del grafico temporale visualizza uno o più insiemi di dati di serie temporali. In genere, si utilizza prima un nodo Intervalli di tempo per creare un campo *EtichettaTempo*, che viene utilizzato per attribuire un'etichetta all'asse x.



Il nodo Valutazione facilita la valutazione e il confronto di modelli predittivi. Il grafico di valutazione mostra il comportamento dei modelli nella previsione di particolari risultati. Ordina i record in base al valore previsto e alla confidenza della previsione, quindi li suddivide in gruppi di uguale dimensione (**quantili**) e infine rappresenta il valore del criterio di business per ciascun quantile, dal più alto al più basso. I modelli multipli sono mostrati nel grafico come linee separate.

Dopo aver aggiunto un nodo Grafico a un flusso, è possibile fare doppio clic su di esso per aprire una finestra di dialogo in cui specificare le opzioni. La maggior parte dei grafici contiene opzioni univoche, presentate in una o più schede. Esistono inoltre diverse opzioni comuni a tutti i grafici. Ulteriori informazioni sulle opzioni comuni sono disponibili nelle sezioni seguenti.

Dopo avere configurato le opzioni relative a un nodo Grafico, è possibile eseguirlo dall'interno della finestra di dialogo o come parte di un flusso. Nella finestra del grafico così generata, è possibile generare nodi Ricava (Insieme e Flag) e nodi Seleziona in base a una selezione o a un'area di dati, dividendo effettivamente i dati in "sottoinsiemi". Questa potente funzione può essere utilizzata, per esempio, per individuare ed escludere i valori anomali.

Effetti estetici, sovrapposizioni, riquadri e animazioni

Sovrapposizioni ed effetti estetici

Gli elementi estetici (e le sovrapposizioni) aggiungono dimensionalità alle visualizzazioni. L'effetto di un elemento estetico (raggruppamento o sovrapposizione) dipende dal tipo di visualizzazione, dal tipo di campo (variabile) e dal tipo e dalle statistiche dell'elemento grafico. Per esempio, per il colore è possibile utilizzare un campo categoriale per raggruppare i punti di un grafico a dispersione o per creare le barre di un grafico a barre sovrapposte. Per il colore è inoltre possibile utilizzare un intervallo numerico continuo per indicare i valori dell'intervallo per ogni punto di un grafico a dispersione.

Per trovare l'elemento estetico e la sovrapposizione adatti alle proprie esigenze, occorre eseguire delle prove. Le seguenti descrizioni possono facilitare la scelta dell'elemento adeguato.

Nota: non tutti gli elementi estetici o le sovrapposizioni sono disponibili per tutti i tipi di visualizzazione.

- **Colore.** Quando il colore è definito da un campo categoriale, suddivide la visualizzazione in base alle singole categorie, un colore per ciascuna categoria. Quando il colore è un intervallo numerico continuo, varia in base al valore del campo intervallo. Se l'elemento grafico (per esempio una barra o una casella) rappresenta più di un record/caso e per il colore viene utilizzato un campo, il colore varia in base alla *media* del campo intervallo.
- **Forma.** La forma è definita da un campo categoriale che suddivide la visualizzazione in elementi di forme diverse, una per ciascuna categoria.
- **Trasparenza.** Quando la trasparenza è definita da un campo categoriale, suddivide la visualizzazione in base alle singole categorie, un livello di trasparenza per ciascuna categoria. Quando la trasparenza è un intervallo numerico continuo, varia in base al valore del campo intervallo. Se l'elemento grafico (per esempio una barra o una casella) rappresenta più di un record/caso e per la trasparenza viene utilizzato un campo intervallo, il colore varia in base alla *media* del campo intervallo. Al valore massimo, gli elementi grafici sono completamente trasparenti, mentre a quello minimo sono completamente opachi.
- **Etichetta dati.** Le etichette dati sono definite da qualsiasi tipo di campo i cui valori sono utilizzati per creare etichette che sono collegate agli elementi grafici.
- **Dimensione.** Quando la dimensione è definita da un campo categoriale, suddivide la visualizzazione in base alle singole categorie, una dimensione per ciascuna categoria. Quando la dimensione è un intervallo numerico continuo, varia in base al valore del campo intervallo. Se l'elemento grafico, per

esempio una barra o una casella, rappresenta più di un record/caso e per la dimensione viene utilizzato un campo intervallo, la dimensione varia in base alla *media* del campo intervallo.

Utilizzo di riquadri e animazioni

Pannelli. La creazione di pannelli, denominati anche facet, crea una tabella di grafici. Per ogni categoria dei campi dei pannelli viene generato un grafico, ma tutti i pannelli vengono visualizzati contemporaneamente. I pannelli sono utili per verificare se la visualizzazione è soggetta alle condizioni dei campi dei pannelli. Per esempio, è possibile creare pannelli in un istogramma in base al sesso per stabilire se le distribuzioni della frequenza sono uguali per maschi e femmine. In altre parole, è possibile verificare se lo stipendio è soggetto a differenze tra i sessi. Selezionare un campo relativo alla categoria per la creazione dei pannelli.

Animazione. L'animazione assomiglia ai pannelli per il fatto che vengono creati più grafici dai valori del campo di animazione, ma tali grafici non vengono mostrati insieme. È invece possibile utilizzare i controlli in modalità Esplora per animare l'output e scorrere una sequenza di singoli grafici. Inoltre, a differenza dei pannelli, l'animazione non richiede un campo categoriale. È possibile specificare un campo continuo i cui valori vengono suddivisi automaticamente in intervalli. È possibile variare la dimensione dell'intervallo con i controlli dell'animazione in modalità Esplora. Non tutte le visualizzazioni offrono la funzionalità di animazione.

Utilizzo della scheda Output

Per tutti i tipi di grafici è possibile specificare le seguenti opzioni relative al nome di file e alla visualizzazione dei grafici generati.

Nota: i grafici del nodo distribuzione dispongono di impostazioni aggiuntive.

Nome output. Specifica il nome del grafico prodotto durante l'esecuzione del nodo. **Automatico** sceglie un nome in base al nodo che genera l'output. Se necessario, è possibile selezionare **Personalizzato** per specificare un nome diverso.

Output su schermo. Selezionare questa opzione per generare e visualizzare il grafico in una nuova finestra.

Output su file. Selezionare questa opzione per salvare l'output come file.

- **Grafico di output.** Selezionare questa opzione per generare l'output in formato grafico. Disponibile solo nei nodi distribuzione.
- **Tabella di output.** Selezionare questa opzione per generare l'output in formato tabella. Disponibile solo nei nodi distribuzione.
- **Nome file.** Specificare un nome di file utilizzato per il grafico o la tabella generati. Utilizzare il pulsante con i puntini di sospensione (...) per specificare un file e una posizione specifici.
- **Tipo di file.** Selezionare il tipo di file dall'elenco a discesa. Per tutti i nodi Grafico, fatta eccezione per il nodo distribuzione con l'opzione **Tabella di output**, i tipi di file disponibili per i grafici sono i seguenti:
 - Bitmap (.bmp)
 - PNG (.png)
 - Oggetto di output (.cou)
 - JPEG (.jpg)
 - HTML (.html)
 - Documento ViZml (.xml) da utilizzare in altre applicazioni IBM SPSS Statistics.

Per l'opzione **Tabella di output** del nodo distribuzione, i tipi di file disponibili sono i seguenti:

- Dati delimitati da tabulazione (.tab)

- Dati delimitati da virgola (.csv)
- HTML (.html)
- Oggetto di output (.cou)

Impagina output. Quando l'output viene salvato come HTML, questa opzione viene abilitata per consentire di controllare la dimensione di ciascuna pagina HTML. Valida solo per il nodo Distribuzione.

Righe per pagina. Quando è selezionata l'opzione **Impagina output**, questa opzione viene abilitata per consentire di determinare la lunghezza di ciascuna pagina HTML. L'impostazione di default è 400 righe. Valida solo per il nodo Distribuzione.

Utilizzo della scheda Annotazioni.

Utilizzata per tutti i nodi, questa scheda fornisce opzioni per rinominare i nodi, uno strumento personalizzato e la possibilità di archiviare annotazioni estese.

Grafici 3-D

In IBM SPSS Modeler, i grafici raccolta e plot consentono di visualizzare le informazioni su un terzo asse. Questa funzione offre all'utente una maggiore flessibilità in fase di visualizzazione dei dati, in quanto consente di selezionare sottoinsiemi o derivare nuovi campi per la creazione di modelli.

Dopo che è stato creato un grafico 3-D, è possibile fare clic su di esso e trascinarlo in modo da ruotarlo e visualizzarlo da tutte le angolature.

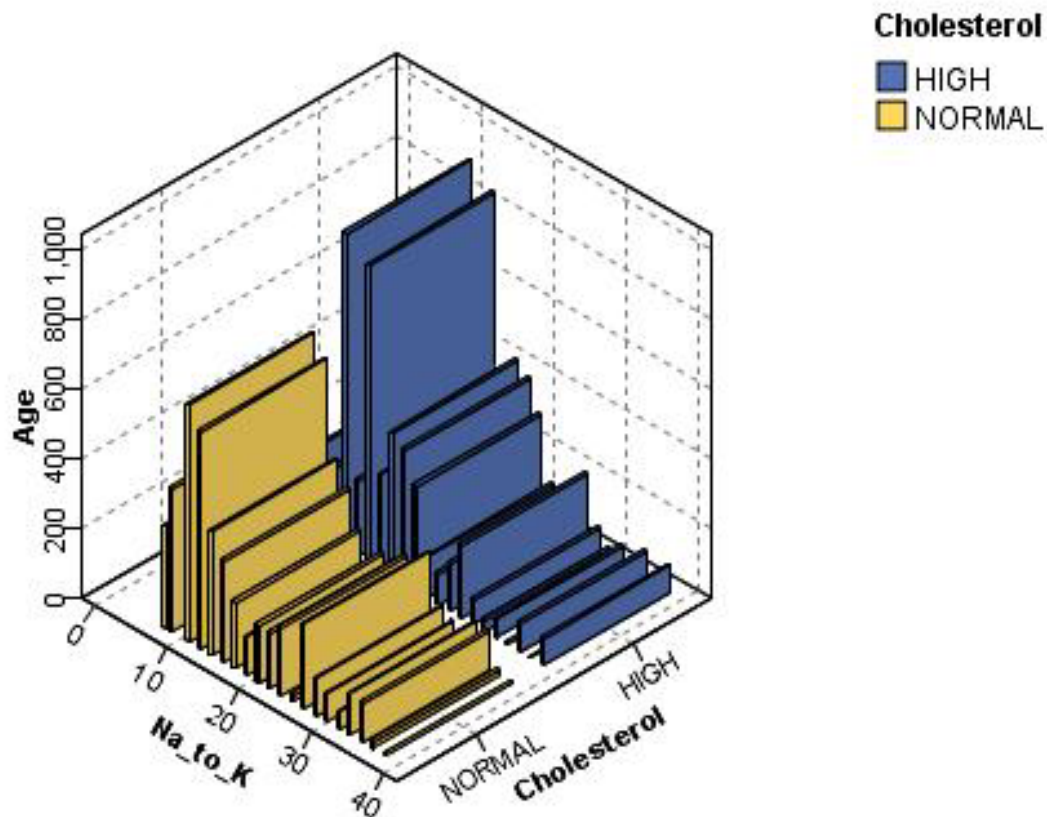


Figura 9. Grafico raccolto con assi x, y e z

In IBM SPSS Modeler esistono due modalità di creazione per i grafici 3-D: tramite rappresentazione delle informazioni su un terzo asse (grafici 3-D effettivi) o tramite visualizzazione di grafici con effetti 3-D. Entrambi i metodi sono disponibili per i grafici e le raccolte.

Per rappresentare le informazioni su un terzo asse

1. Nella finestra di dialogo del nodo Grafico, fare clic sulla scheda **Plot**.
2. Fare clic sul pulsante 3D per attivare le opzioni per l'asse z.
3. Utilizzare il pulsante Selettore di campo per selezionare un campo per l'asse z. In alcuni casi, per l'asse z è possibile utilizzare solo campi simbolici. Selettore di campo visualizzerà i campi appropriati.

Per aggiungere effetti 3-D a un grafico

1. Dopo aver creato un grafico, fare clic sulla scheda **Grafico** nella finestra di output.
2. Fare clic sul pulsante 3-D per visualizzare un grafico tridimensionale.

Nodo Lavagna grafica

Il nodo Lavagna grafica consente di scegliere tra numerosi output di grafico differenti (grafici a barre, a torta, istogrammi, a dispersione, mappe di calore e così via) in un singolo nodo. Per iniziare, scegliere i campi di dati da analizzare nella prima scheda. Il nodo visualizza un elenco di tipi di grafici adatti ai dati prescelti. Il nodo esclude automaticamente tutti i tipi di grafici non adatti ai campi selezionati. È possibile definire opzioni dettagliate o più avanzate per i grafici nella scheda Dettagliato.

Nota: per modificare il nodo Lavagna grafica o selezionare dei tipi di grafici è necessario collegare il nodo Lavagna grafica a un flusso contenente dei dati.

Sono disponibili due pulsanti che consentono di controllare i modelli di visualizzazione (e fogli di stile e mappe) disponibili:

Gestisci. Gestire modelli, fogli di stile e mappe di visualizzazione sul computer. È possibile importare, esportare, rinominare ed eliminare modelli, fogli di stile e mappe di visualizzazione sul computer locale. Per ulteriori informazioni, consultare l'argomento "Gestione dei modelli, dei fogli di stile e dei file delle mappe" a pagina 196.

Posizione. Modificare la posizione in cui sono archiviati i modelli, i fogli di stile e le mappe di visualizzazione. La posizione corrente viene elencata alla destra del pulsante. Per ulteriori informazioni, consultare l'argomento "Impostazione della posizione di modelli, fogli di stile e mappe" a pagina 196.

Scheda Di base del nodo Lavagna grafica

Se non si è certi di quale sia il tipo di visualizzazione ottimale per rappresentare i propri dati, utilizzare la scheda Di base. Quando si selezionano i dati, viene visualizzato un sottoinsieme dei tipi di visualizzazione adatti per rappresentarli. Per alcuni esempi, consultare "Lavagna grafica Esempi" a pagina 184.

1. Selezionare uno o più campi (variabili) nell'elenco. Per selezionare più campi, tenere premuto il tasto Ctrl e fare clic con il mouse.
Si noti che il livello di misurazione del campo determina i tipi di visualizzazione disponibili. Per cambiare il livello di misurazione, fare clic con il pulsante destro del mouse sul campo nell'elenco e scegliere un'opzione. Per ulteriori informazioni sui tipi di livelli di misurazione disponibili, consultare "Tipi di campo (variabile)" a pagina 172.
2. Selezionare un tipo di visualizzazione. Per le descrizioni dei tipi disponibili, consultare "Tipi di visualizzazione della lavagna grafica incorporati disponibili" a pagina 176.
3. Per alcune visualizzazioni è possibile selezionare una statistica riassuntiva. Sono disponibili diversi sottoinsiemi di statistiche a seconda del fatto che la statistica sia basata sul conteggio o calcolata a partire da un campo continuo. Le statistiche disponibili dipendono anche dal modello. La fase successiva è seguita da un elenco completo delle statistiche che potrebbero essere disponibili.
4. Se si desidera definire altre opzioni, quali gli elementi estetici facoltativi e i campi del pannello, fare clic su **Dettagliato**. Per ulteriori informazioni, consultare l'argomento "Lavagna grafica Scheda Dettagliato" a pagina 174.

Statistiche riassuntive calcolate a partire da un campo continuo

- *Media.* Una misura di tendenza centrale. Media aritmetica, ovvero somma divisa per il numero di casi.
- *Median.* È il valore sopra il quale e sotto il quale ricade la metà dei casi, il 50-esimo percentile. Se il numero di casi è pari, la mediana è pari alla media dei due casi centrali quando questi sono ordinati secondo l'ordine ascendente o discendente. La mediana è una misura di tendenza centrale non sensibile ai valori anomali, a differenza della media che può essere influenzata da valori eccezionalmente bassi o alti.
- *Modalità.* Il valore che ricorre più frequentemente. Se più valori condividono la maggiore ricorrenza, ognuno di essi è una modalità.

- *Minimo*. Il valore più basso di una variabile numerica.
- *Massimo*. Il valore più alto di una variabile numerica.
- *Intervallo*. Differenza tra i valori minimi e massimi.
- *Intervallo medio*. Il punto medio dell'intervallo, vale a dire, il valore la cui differenza rispetto al punto minimo è uguale alla differenza rispetto al punto massimo.
- *Somma*. La somma o il totale di tutti i valori non mancanti di tutti i casi.
- *Somma cumulata*. La somma cumulata dei valori. Ogni elemento grafico mostra la somma di un sottogruppo più la somma totale di tutti i gruppi precedenti.
- *Somma percentuale*. La percentuale all'interno di ogni sottogruppo basata su un campo sommato confrontato alla somma di tutti i gruppi.
- *Somma percentuale cumulata*. La percentuale cumulativa all'interno di ogni sottogruppo basata su un campo sommato confrontato alla somma di tutti i gruppi. Ogni elemento grafico mostra la percentuale di un sottogruppo più la percentuale totale di tutti i gruppi precedenti.
- *Varianza*. Una misura della dispersione dei valori intorno alla media. È calcolata come somma dei quadrati degli scostamenti dalla media, divisa per un valore corrispondente alla somma dei casi meno uno. La varianza è espressa in quadrati dell'unità di misura della variabile.
- *Deviazione standard*. Una misura di dispersione intorno alla media. In una distribuzione normale, il 68% dei casi rientra in una deviazione standard della media e il 95% dei casi rientra in due deviazioni standard. Se, ad esempio, in una popolazione con distribuzione normale l'età media fosse 45 e la deviazione standard 10, il 95% dei casi cadrebbe fra 25 e 65 anni.
- *Errore standard*. Una misura di quanto il valore di una statistica del test può variare da campione a campione. È la deviazione standard della distribuzione di campionamento di una statistica. L'errore standard della media, ad esempio, è la deviazione standard delle medie del campione.
- *Curtosi*. Una misura di quanto le osservazioni si trovino raggruppate in un punto centrale. Per la distribuzione normale, il valore della statistica di curtosi è zero. Una curtosi positiva indica che, rispetto a una distribuzione normale, le osservazioni sono più raggruppate intorno al centro della distribuzione e hanno code più sottili fino ai valori estremi della distribuzione; a quel punto, le code della distribuzione leptocurtica sono più spesse rispetto a una distribuzione normale. Una curtosi negativa indica che, rispetto a una distribuzione normale, le osservazioni sono meno raggruppate e hanno code più spesse fino ai valori estremi della distribuzione; a quel punto, le code della distribuzione platicurtica sono più sottili rispetto a una distribuzione normale.
- *Asimmetria*. Una misura dell'asimmetria di una distribuzione. La distribuzione normale è simmetrica e ha un valore di asimmetria pari a 0. Una distribuzione con una notevole asimmetria positiva ha una lunga coda a destra. Una distribuzione con asimmetria negativa ha una coda a sinistra. In generale un'asimmetria con valore più che doppio dell'errore standard indica lo scostamento dalla normale simmetria.

Le seguenti statistiche di regione potrebbero restituire più elementi grafici per ogni sottogruppo. Quando si utilizzano gli elementi grafici intervallo, area o margine, una statistica di regione restituisce un elemento grafico che mostra l'intervallo. Tutti gli altri elementi grafici danno come risultato due elementi separati, uno che mostra l'inizio dell'intervallo e l'altro che mostra la fine dell'intervallo.

- **Regione: Intervallo**. L'intervallo di valori tra il valore minimo e il valore massimo.
- **Regione: Intervallo di confidenza per la media al 95%**. Un intervallo di valori con il 95% di probabilità di includere la media della popolazione.
- **Regione: Intervallo di confidenza per il singolo al 95%**. Un intervallo di valori con il 95% di probabilità di includere il valore previsto dato il caso singolo.
- **Regione: 1 Deviazione standard sopra/sotto la media**. Un intervallo di valori compreso tra 1 *deviazione standard* sopra e sotto la *media*.
- **Regione: 1 Errore standard sopra/sotto la media**. Un intervallo di valori compreso tra 1 *errore standard* sopra e sotto la *media*.

Statistiche riassuntive basate sul conteggio

- **Conteggio.** Il numero di righe/casi.
- **Conteggio cumulato.** Il numero cumulato di righe/casi. Ogni elemento grafico mostra il conteggio riferito a un sottogruppo più il conteggio totale di tutti i gruppi precedenti.
- **Percentuale del conteggio.** La percentuale di righe/casi in ogni sottogruppo rispetto al numero totale di righe/casi.
- **Percentuale cumulata del conteggio.** La percentuale cumulata di righe/casi in ogni sottogruppo rispetto al numero totale di righe/casi. Ogni elemento grafico mostra la percentuale di un sottogruppo più la percentuale totale di tutti i gruppi precedenti.

Tipi di campo (variabile)

Negli elenchi di campi, accanto ai campi appaiono delle icone che indicano il tipo di campo e il tipo di dati. Le icone identificano anche gli insiemi di risposte multiple.

Tabella 42. Icone livello di misurazione.

Livello di misurazione	Numeriche	Stringa	Data	Ora
Continuo		n/d		
Insieme ordinato				
Set				

Tabella 43. Icone insieme di risposte multiple.

Tipo insieme di risposte multiple	Icona
Insieme di risposte multiple, categorie multiple	
Insieme di risposte multiple, dicotomie multiple	

Livello di misurazione

Il livello di misurazione di un campo è importante quando si crea una visualizzazione. Di seguito viene fornita una descrizione dei livelli di misurazione. Per cambiare temporaneamente un livello di misurazione, fare clic con il pulsante destro del mouse su un campo in un elenco di campi e scegliere un'opzione. Nella maggior parte dei casi, è necessario tenere in considerazione solo le due classificazioni più ampie di campi, vale a dire, categoriali e continui:

Categoriali. Dati con un numero limitato di categorie o valori distinti (per esempio, genere o religione). I campi categoriali possono essere stringhe (alfanumerici) o campi numerici che utilizzano codici numerici per rappresentare categorie (per esempio, 0 = *maschio* e 1 = *femmina*). Sono definiti anche dati qualitativi. Insiemi, insiemi ordinati e flag sono tutti campi categoriali.

- **Insieme.** Un campo/variabile i cui valori rappresentano le categorie senza ordinamento intrinseco (ad esempio, l'ufficio dell'azienda un impiegato). Tra gli esempi di variabili nominali troviamo la regione, il codice postale e la religione. Noto anche come variabile nominale.
- **Insieme ordinato.** Un campo/variabile i cui valori rappresentano le categorie con qualche ordinamento intrinseco (ad esempio, i livelli di soddisfazione del servizio, da molto insoddisfatto a molto

soddisfatto). Alcuni esempi di insiemi ordinati includono i punteggi di atteggiamento che rappresentano i gradi di soddisfazione o fiducia ed i punteggi di preferenza. Noto anche come variabile ordinale.

- *Flag*. Un campo/variabile con due valori distinti, come Sì o No oppure 1 e 2. Noto anche come variabile dicotomica o binaria.

Continuo. Dati misurati su una scala di intervalli o rapporti, dove i valori dei dati indicano sia l'ordine dei valori che la distanza tra i valori. Per esempio, uno stipendio di € 72.195 è superiore a uno stipendio di € 52.398 e la distanza tra i due valori è € 19.797. Sono definiti anche dati quantitativi, di scala o di intervallo numerico.

I campi categoriali definiscono le categorie nella visualizzazione, tipicamente per creare elementi grafici distinti o per raggruppare gli elementi grafici. I campi continui sono spesso riepilogati all'interno di categorie di campi categoriali. Per esempio, in una visualizzazione predefinita relativa al reddito per le categorie Sesso verranno visualizzati il reddito medio dei maschi e il reddito medio delle femmine. I valori grezzi relativi ai campi continui possono anche essere rappresentati graficamente, come nel caso dei grafici a dispersione. Per esempio, un grafico a dispersione potrebbe mostrare lo stipendio attuale e lo stipendio iniziale per ogni caso. Un campo categoriale potrebbe essere usato per raggruppare i casi in base al sesso.

Tipi di dati

Il livello di misurazione non è l'unica proprietà che determina il tipo di un campo. Il campo viene anche memorizzato come tipo di dati specifico, per esempio, stringhe (dati non numerici, come le lettere), valori numerici (numeri reali) e date. A differenza del livello di misurazione, il tipo di dati di un campo non può essere temporaneamente modificato. È necessario modificare il modo in cui i dati sono memorizzati nell'insieme di dati originale.

Insiemi a risposta multipla

Alcuni file di dati supportano un tipo di "campo" speciale denominato **insieme a risposta multipla**. Gli insiemi di risposte multiple non sono in realtà "campi" in senso normale. Gli insiemi di risposte multiple utilizzano campi multipli per registrare le risposte alle domande per cui l'intervistato può fornire più di una risposta. Vengono trattati come campi categoriali e le operazioni supportate dai campi categoriali possono, per la maggior parte, essere eseguite anche con gli insiemi di risposte multiple.

Gli insiemi di risposte multiple possono essere insiemi dicotomici multipli o insiemi di categorie multiple.

Insiemi a dicotomie multiple. Un insieme a dicotomia multipla generalmente è composto da più campi dicotomici: campi con solo due valori possibili come sì/no, presente/assente, selezionato/non selezionato. Anche se i campi potrebbero non essere tutti rigorosamente dicotomici, tutti i campi nell'insieme sono codificati allo stesso modo.

Per esempio, si immagina un'indagine che offre cinque risposte possibili alla domanda, "Quale delle seguenti fonti di informazioni utilizzi per tenerti informato"? Il rispondente può indicare più risposte selezionando la casella accanto a ogni opzione. Le cinque risposte diventano cinque campi nel file di dati, codificate con 0 per *No* (non selezionata) e 1 per *Sì* (selezionata).

Insiemi di categorie multiple. Un insieme di categorie multiple è formato da più campi, tutti codificati allo stesso modo, spesso con molte categorie di risposta possibili. Per esempio, si supponga la seguente domanda in un'indagine "Indica un massimo di tre nazionalità che meglio descrivono la tua appartenenza etnica". Potrebbero esserci centinaia di risposte possibili ma, per esigenze di codificazione, l'elenco è limitato alle 40 nazionalità più comuni, mentre tutte le altre possibilità sono incluse nella categoria "altro". Nel file di dati, le tre scelte diventeranno tre campi, ognuno con 41 categorie (40 nazionalità codificate e una categoria "altro").

Lavagna grafica Scheda Dettagliato

Utilizzare la scheda Dettagliato quando si conosce il tipo di visualizzazione da creare o quando si desidera aggiungere elementi estetici facoltativi, pannelli e/o effetti di animazione a una visualizzazione. Per alcuni esempi, consultare “Lavagna grafica Esempi” a pagina 184.

1. Verrà visualizzato il tipo di visualizzazione scelto nella scheda Di base, se ne è stato selezionato uno; in caso contrario, sceglierne uno dall'elenco a discesa. Per informazioni relative ai tipi di visualizzazione, consultare “Tipi di visualizzazione della lavagna grafica incorporati disponibili” a pagina 176.
2. Subito a destra della miniatura della visualizzazione si trovano i comandi per specificare i campi (le variabili) necessari per il tipo di visualizzazione. È necessario specificare tutti questi campi.
3. Per alcune visualizzazioni è possibile selezionare una statistica riassuntiva. In taluni casi, per esempio nei grafici a barre, è possibile utilizzare una di queste opzioni riassuntive per l'effetto estetico della trasparenza. Per le descrizioni delle statistiche riassuntive, consultare “Scheda Di base del nodo Lavagna grafica” a pagina 170.
4. È possibile selezionare uno o più degli elementi estetici facoltativi. Essi aggiungono dimensionalità consentendo l'inclusione di altri campi nella visualizzazione. Per esempio, è possibile utilizzare un campo per variare la dimensione dei punti in un grafico a dispersione. Per ulteriori informazioni relative agli elementi estetici facoltativi, consultare “Effetti estetici, sovrapposizioni, riquadri e animazioni” a pagina 166. Si noti che l'effetto estetico della trasparenza non è supportato tramite gli script.
5. Se si sta creando una visualizzazione delle mappe, il gruppo **File mappa** mostra il file o i file delle mappe che saranno utilizzati. Viene visualizzato l'eventuale file delle mappe di default. Per modificare il file delle mappe, fare clic su **Seleziona un file mappa** per visualizzare la finestra di dialogo Seleziona mappe. In questa finestra di dialogo è possibile anche specificare il file delle mappe di default. Per ulteriori informazioni, consultare l'argomento “Selezione dei file delle mappe per le visualizzazioni delle mappe”.
6. È possibile selezionare una o più opzioni relative ai pannelli o all'animazione. Per ulteriori informazioni sulle opzioni relative ai pannelli e all'animazione, vedere “Effetti estetici, sovrapposizioni, riquadri e animazioni” a pagina 166.

Selezione dei file delle mappe per le visualizzazioni delle mappe

Se si seleziona un modello di visualizzazione delle mappe, è necessario un file delle mappe che definisca le informazioni geografiche per disegnare la mappa. Se è presente un file delle mappe di default, viene utilizzato per la visualizzazione delle mappe. Per scegliere un file delle mappe differente, fare clic su **Seleziona un file mappa** nella scheda Dettagliato per visualizzare la finestra di dialogo Seleziona mappe.

La finestra di dialogo Seleziona mappe consente di selezionare un file delle mappe principale e un file delle mappe di riferimento. I file delle mappe definiscono le informazioni geografiche per disegnare le mappe. L'applicazione è installata con diversi file delle mappe standard. Se si dispone di altri shapefile ESRI che si desidera utilizzare, è necessario convertire prima gli shapefile in file SMZ. Per ulteriori informazioni, consultare l'argomento “Conversione e distribuzione degli shapefile delle mappe” a pagina 197. Una volta convertita la mappa, fare clic su **Gestisci...** nella finestra di dialogo Selezionatore modelli per importare la mappa nel sistema di gestione, in modo che sia disponibile nella finestra di dialogo Seleziona mappe.

Si riportano di seguito alcuni punti da considerare quando si specificano file delle mappe:

- Tutti i modelli delle mappe richiedono almeno un file delle mappe.
- Il file delle mappe collega generalmente un attributo chiave della mappa a una chiave di dati.
- Se il modello non necessita di una chiave della mappa che si colleghi a una chiave dei dati, necessita di un file e dei campi delle mappe di riferimento che specifichino le coordinate (ad esempio longitudine e latitudine) per disegnare gli elementi nella mappa di riferimento.

- La sovrapposizione dei modelli di mappa richiede due mappe: un file delle mappe principale ed un file delle mappe di riferimento. Il file delle mappe di riferimento viene disegnato per primo in modo che si trovi dietro al file delle mappe principale.

Per informazioni sulla terminologia relativa alle mappe, come attributi e funzioni, consultare “Concetti principali delle mappe” a pagina 198.

File mappa. È possibile selezionare qualunque file delle mappe presente nel sistema di gestione. I file comprendono i file delle mappe preinstallati che sono stati importati. Per ulteriori informazioni relative alla gestione dei file mappa, consultare “Gestione dei modelli, dei fogli di stile e dei file delle mappe” a pagina 196.

Chiave mappa. Specificare l'attributo che si vuole utilizzare come chiave che collega il file delle mappe alla chiave dei dati.

Salva questo file mappa e le impostazioni come valori predefiniti. Selezionare questa casella di controllo se si vuole utilizzare il file delle mappe selezionato come default. Se si è specificato un file delle mappe di default, non è necessario specificare un file delle mappe ogni volta che si crea una visualizzazione delle mappe.

Chiave dati. Questo comando elenca gli stessi valori che appaiono nella scheda Dettagliato del Selezionatore modelli. È presentato qui per comodità se si desidera cambiare la chiave in funzione dello specifico file delle mappe che si è scelto.

Visualizza tutte le caratteristiche della mappa nella visualizzazione. Quando questa opzione è selezionata, si interpretano tutte le funzioni della mappa nella visualizzazione, anche se non vi sono valori corrispondenti nella chiave di dati. Se si desidera visualizzare solo le funzioni per cui sono disponibili dati, annullare la selezione di questa opzione. Nella visualizzazione non verranno interpretate le caratteristiche identificate dalle chiavi della mappa presenti nell'elenco **Chiavi mappa senza corrispondenza**.

Confronta la mappa e i valori dei dati. La chiave della mappa e la chiave dei dati si collegano tra loro per creare una visualizzazione della mappa. La chiave mappa e la chiave di dati devono essere estratte dallo stesso dominio (ad esempio, paesi e regioni). Fare clic su **Confronta** per verificare se i valori della chiave dei dati e della chiave della mappa corrispondono. L'icona visualizzata informa sullo stato del confronto. Le icone sono descritte di seguito. Se è stato effettuato un confronto e vi sono chiavi dei dati senza corrispondenti chiavi della mappa, i valori delle chiavi dei dati sono mostrati nell'elenco **Chiavi dati senza corrispondenza**. Nell'elenco **Chiavi mappa senza corrispondenza**, è anche possibile vedere quali valori delle chiavi della mappa non hanno valori delle chiavi dati corrispondenti. Se l'opzione **Visualizza tutte le caratteristiche della mappa nella visualizzazione** non è selezionata, le funzioni identificate da questi valori della chiave della mappa non verranno visualizzati.

Tabella 44. Icone di confronto.





Icona	Descrizione
	Non è stato eseguito alcun confronto. Questo è lo stato di default prima di fare clic su Confronta . Si consiglia di procedere con cautela perché non è noto se i valori delle chiavi dei dati e delle chiavi delle mappe corrispondono.
	È stato effettuato un confronto, e i valori della chiave dei dati e della chiave della mappa corrispondono. Per ogni valore della chiave dei dati è stata individuata una caratteristica corrispondente nella chiave della mappa.

Tabella 44. Icone di confronto (Continua).

Icona	Descrizione
	È stato effettuato un confronto, e alcuni valori della chiave dei dati e della chiave della mappa non corrispondono. Per alcuni valori della chiave dei dati non è stata individuata una caratteristica corrispondente nella chiave della mappa. Procedere con attenzione. Se si prosegue, la visualizzazione della mappa non conterrà tutti i valori dei dati.
	È stato effettuato un confronto, e nessuno dei valori della chiave dei dati e della chiave della mappa corrisponde. Si consiglia di selezionare una chiave dei dati o una chiave della mappa diversa perché se si prosegue non sarà interpretata alcuna mappa.

Tipi di visualizzazione della lavagna grafica incorporati disponibili

È possibile creare diversi tipi di visualizzazione differenti. Tutti i tipi incorporati riportati di seguito sono disponibili nelle schede Di base e Dettagliato. Alcune delle descrizioni per i modelli (in particolare i modelli di mappa) identificano i campi (variabili) specificati nella scheda Dettagliato utilizzando **testo speciale**.

Tabella 45. Tipi di grafici disponibili.

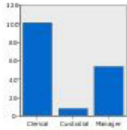
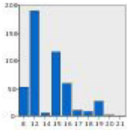
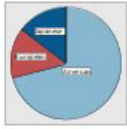
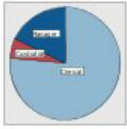
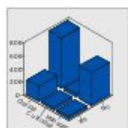
Icona Grafico	Descrizione	Icona Grafico	Descrizione
	<p>Barre</p> <p>Calcola una statistica di riepilogo per un campo numerico continuo e visualizza i risultati per ciascuna categoria di un campo relativo alla categoria come barre.</p> <p><i>Richiede:</i> un campo relativo alla categoria ed un campo continuo.</p>		<p>Barre dei conteggi</p> <p>Visualizza la proporzione di righe/casi in ciascuna categoria di un campo relativo alla categoria come barre. Per generare questo grafico è anche possibile utilizzare il nodo grafico Distribuzione che dispone di alcune opzioni aggiuntive. Per ulteriori informazioni, consultare l'argomento "Nodo distribuzione" a pagina 211.</p> <p><i>Richiede:</i> un singolo campo relativo alla categoria.</p>
	<p>Torta</p> <p>Calcola la somma di un campo numerico continuo e visualizza la proporzione di tale somma distribuita in ciascuna categoria di un campo relativo alla categoria come fette di una torta.</p> <p><i>Richiede:</i> un campo relativo alla categoria ed un campo continuo.</p>		<p>Torta dei conteggi</p> <p>Visualizza la proporzione di righe/casi in ciascuna categoria di un campo relativo alla categoria come fette di una torta.</p> <p><i>Richiede:</i> un singolo campo relativo alla categoria.</p>

Tabella 45. Tipi di grafici disponibili (Continua).

Icona Grafico



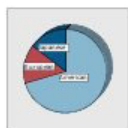
Descrizione

Barre 3D

Calcola una statistica di riepilogo per un campo numerico continuo e visualizza i risultati dell'incrocio di categorie tra due campi relativi alla categoria.

Richiede: una coppia di campi relativi alla categoria ed un campo continuo.

Icona Grafico

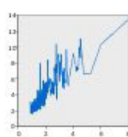


Descrizione

Torta 3D

È uguale al grafico a torta, con l'aggiunta di un effetto tridimensionale.

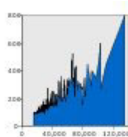
Richiede: un campo relativo alla categoria ed un campo continuo.



Line

Calcola una statistica di riepilogo per ciascun valore di un altro campo e traccia una linea che collega i valori. È anche possibile generare un grafico a linee mediante il nodo grafico che dispone di alcune opzioni aggiuntive. Per ulteriori informazioni, consultare l'argomento "Nodo Plot" a pagina 205.

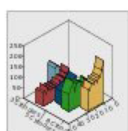
Richiede: una coppia di campi di qualsiasi tipo.



Area

Calcola una statistica di riepilogo per un campo per ciascun valore di un altro campo e visualizza un'area che collega i valori. La differenza tra una linea e un'area è minima, poiché un'area è simile a una linea con lo spazio colorato al di sotto di essa. Tuttavia, se si utilizza un elemento estetico colore, si ottiene una semplice suddivisione della linea e una sovrapposizione dell'area.

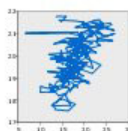
Richiede: una coppia di campi di qualsiasi tipo.



Area 3D

Visualizza i valori di un campo tracciati rispetto ai valori di un altro campo e suddivisi mediante un campo relativo alla categoria. Per ogni categoria viene tracciato un elemento area.

Richiede: un campo relativo alla categoria ed un paio di campi di qualsiasi tipo.



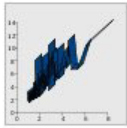
Percorso

Visualizza i valori di un campo tracciati in base ai valori di un altro campo, con una linea che collega i valori nell'ordine in cui sono visualizzati nell'insieme di dati originale. L'ordinamento è la differenza principale tra un percorso e una linea.

Richiede: una coppia di campi di qualsiasi tipo.

Tabella 45. Tipi di grafici disponibili (Continua).

Icona Grafico



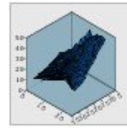
Descrizione

Nastro

Calcola una statistica di riepilogo per un campo per ciascun valore di un altro campo e traccia un nastro che collega i valori. Un nastro è sostanzialmente una linea con effetti tridimensionali, non un vero e proprio grafico 3D.

Richiede: una coppia di campi di qualsiasi tipo.

Icona Grafico



Descrizione

Superficie

Visualizza i valori di tre campi tracciati rispetto ai valori di un altro campo, con una superficie che collega i valori.

Richiede : tre campi di qualsiasi tipo.

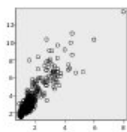


Grafico a dispersione

Visualizza i valori di un campo tracciati rispetto ai valori di un altro campo. Questo grafico può evidenziare la relazione (se presente) tra i campi. Per generare un grafico a dispersione è anche possibile utilizzare il nodo grafico che dispone di alcune opzioni aggiuntive. Per ulteriori informazioni, consultare l'argomento "Nodo Plot" a pagina 205.

Richiede: una coppia di campi di qualsiasi tipo.

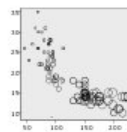


Grafico a bolle

Come un grafico a dispersione di base, visualizza i valori di un campo tracciati rispetto ai valori di un altro campo. La differenza è che vengono utilizzati i valori di un terzo campo per modificare la dimensione dei singoli punti.

Richiede : tre campi di qualsiasi tipo.

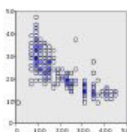


Grafico a dispersione in bin

Come un grafico a dispersione di base, visualizza i valori di un campo tracciati rispetto ai valori di un altro campo. La differenza è data dal fatto che valori simili vengono discretizzati in gruppi e che viene utilizzato l'elemento estetico colore o dimensioni per indicare il numero di casi in ogni bin.

Richiede: una coppia di campi continui.

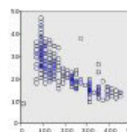


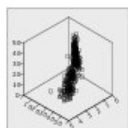
Grafico a dispersione in bin esagonali

Vedere la descrizione relativa al grafico a dispersione in bin. La differenza è data dalla forma dei bin sottostanti, che vengono visualizzati come esagoni anziché come cerchi. Il grafico a dispersione in bin esagonali risultante è simile al grafico a dispersione in bin. Tuttavia, nei due grafici il numero di valori di ogni bin è diverso, per via della forma dei bin sottostanti.

Richiede: una coppia di campi continui.

Tabella 45. Tipi di grafici disponibili (Continua).

Icona Grafico



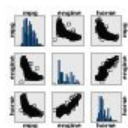
Descrizione

Grafico a dispersione 3D

Visualizzai valori dei tre campi tracciati l'uno in relazione all'altro. Questo grafico può evidenziare la relazione (se presente) tra i campi. Per generare un grafico a dispersione 3D è anche possibile utilizzare il nodo grafico che dispone di alcune opzioni aggiuntive. Per ulteriori informazioni, consultare l'argomento "Nodo Plot" a pagina 205.

Richiede : tre campi di qualsiasi tipo.

Icona Grafico

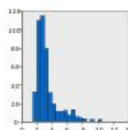


Descrizione

matrice dei grafici a dispersione (SPLOM)

Visualizza i valori di un campo tracciati rispetto ai valori di un altro per ciascun campo. Un grafico SPLOM è come una tabella di grafici a dispersione. Il grafico SPLOM include anche un istogramma di ciascun campo.

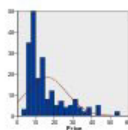
Richiede: due o più campi continui.



Istogramma

Visualizza la distribuzione della frequenza di un campo. Un istogramma può essere utile per determinare il tipo di distribuzione e per rilevare l'eventuale asimmetria della distribuzione. Per generare questo grafico è anche possibile utilizzare il nodo grafico Istogramma che dispone di alcune opzioni aggiuntive. Per ulteriori informazioni, consultare l'argomento "Scheda Plot del nodo Istogramma" a pagina 215.

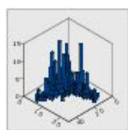
Richiede: un singolo campo di qualsiasi tipo.



Istogramma con distribuzione normale

Visualizza la distribuzione della frequenza di un campo continuo con una curva della distribuzione normale sovrapposta.

Richiede: un campo continuo singolo.



Istogramma 3D

Visualizza la distribuzione della frequenza di una coppia di campi continui.

Richiede: una coppia di campi continui.



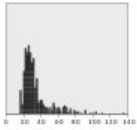
Densità 3D

Visualizza la distribuzione della frequenza di una coppia di campi continui. È simile all'istogramma 3D, con la sola differenza che per visualizzare la distribuzione viene utilizzata una superficie anziché delle barre.

Richiede: una coppia di campi continui.

Tabella 45. Tipi di grafici disponibili (Continua).

Icona Grafico



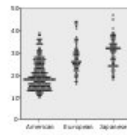
Descrizione

Grafico a punti

Riproduce i singoli casi/righe e li sovrappone in punti dati separati sull'asse x. Il grafico è simile a un istogramma poiché mostra la distribuzione dei dati, ma visualizza i singoli casi/righe anziché un conteggio aggregato di un bin specifico (intervallo di valori).

Richiede: un singolo campo di qualsiasi tipo.

Icona Grafico



Descrizione

Grafico a punti 2D

Visualizza i singoli casi/righe e li sovrappone in punti dati distinti sull'asse y per ciascuna categoria di un campo relativo alla categoria.

Richiede: un campo relativo alla categoria ed un campo continuo.

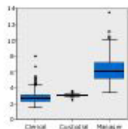
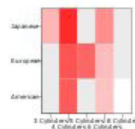


Grafico a scatole

Calcola le cinque statistiche (minimo, primo quartile, mediana, terzo quartile e massimo) per un campo continuo per ciascuna categoria di un campo relativo alla categoria. I risultati vengono visualizzati in forma di grafico a scatole/schema. I grafici a scatole sono utili per vedere in che modo la distribuzione dei dati continui varia tra le categorie.

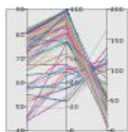
Richiede: un campo relativo alla categoria ed un campo continuo.



Mappa termica

Calcola la media per un campo continuo per l'incrocio delle categorie tra due campi relativi alla categoria.

Richiede: una coppia di campi relativi alla categoria ed un campo continuo.



Parallelo

Crea assi paralleli per ciascun campo e traccia una linea attraverso il valore del campo per ciascuna riga/caso nei dati.

Richiede: due o più campi continui.



Coropleta di conteggi

Calcola il conteggio per ciascuna categoria di un campo relativo alla categoria (**Chiave dati**) e traccia una mappa che utilizza la saturazione dei colori per rappresentare i conteggi nelle funzioni della mappa che corrispondono alle categorie.

Richiede: un campo relativo alla categoria. Un file di mappa la cui chiave corrisponde alle categorie **Chiave dati**.

Tabella 45. Tipi di grafici disponibili (Continua).

Icona Grafico



Descrizione

Coropleta di medie/mediane/somme

Calcola la media, la mediana o la somma di un campo continuo (**Colore**) per ciascuna categoria di un campo relativo alla categoria (**Chiave dati**) e traccia una mappa che utilizza la saturazione dei colori per rappresentare le statistiche calcolate nelle funzioni della mappa che corrispondono alle categorie.

Richiede: un campo relativo alla categoria ed un campo continuo. Un file di mappa la cui chiave corrisponde alle categorie **Chiave dati**.

Icona Grafico



Descrizione

Coropleta di valori

Traccia una mappa che utilizza i colori per rappresentare i valori di un campo relativo alla categoria (**Colore**) per le funzioni della mappa che corrispondono ai valori definiti da un altro campo relativo alla categoria (**Chiave dati**). Se sono presenti più valori relativi alla categoria del campo Colore per ciascuna funzione, viene utilizzato il valore modale.

Richiede: una coppia di campi relativi alla categoria. Un file di mappa la cui chiave corrisponde alle categorie **Chiave dati**.



Coordinate su una coropleta di conteggi

È simile a Coropleta di conteggi, con l'eccezione che sono presenti due ulteriori campi continui (**Longitudine** e **Latitudine**) che identificano le coordinate per la traccia dei punti sulla coropleta.

Richiede: un campo relativo alla categoria ed una coppia di campi continui. Un file di mappa la cui chiave corrisponde alle categorie **Chiave dati**.



Coordinate su una coropleta di medie/mediane/somme

È simile a Coropleta di medie/mediane/somme, con l'eccezione che sono presenti due ulteriori campi continui (**Longitudine** e **Latitudine**) che identificano le coordinate per la traccia dei punti sulla coropleta.

Richiede: un campo relativo alla categoria e tre campi continui. Un file di mappa la cui chiave corrisponde alle categorie **Chiave dati**.



Coordinate su una coropleta di valori

È simile a Coropleta di valori con l'eccezione che sono presenti due ulteriori campi continui (**Longitudine** e **Latitudine**) che identificano le coordinate per la traccia dei punti sulla coropleta.

Richiede: una coppia di campi relativi alla categoria ed una coppia di campi continui. Un file di mappa la cui chiave corrisponde alle categorie **Chiave dati**.



Barre di conteggi su una mappa

Calcola la proporzione di righe/casi in ciascuna categoria di un campo relativo alla categoria (**Categorie**) per ciascuna funzione della mappa (**Chiave dati**) e traccia una mappa ed un grafico a barre al centro di ciascuna funzione della mappa.

Richiede: una coppia di campi relativi alla categoria. Un file di mappa la cui chiave corrisponde alle categorie **Chiave dati**.

Tabella 45. Tipi di grafici disponibili (Continua).

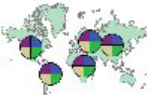


Icona Grafico	Descrizione	Icona Grafico	Descrizione
	<p>Barre su una mappa</p> <p>Calcola una statistica di riepilogo per un campo continuo (Valori) e visualizza i risultati per ciascuna categoria di un campo relativo alla categoria (Categorie) per ciascuna funzione della mappa (Chiave dati) come grafici a barre posizionati al centro di ciascuna funzione della mappa.</p> <p><i>Richiede:</i> una coppia di campi relativi alla categoria ed un campo continuo. Un file di mappa la cui chiave corrisponde alle categorie Chiave dati.</p>		<p>Torta di conteggi di una mappa</p> <p>Visualizza la proporzione di righe/casi in ciascuna categoria di un campo relativo alla categoria (Categorie) per ciascuna funzione della mappa (Chiave dati) e traccia una mappa e le proporzioni come fette di un grafico a torta al centro di ciascuna funzione della mappa.</p> <p><i>Richiede:</i> una coppia di campi relativi alla categoria. Un file di mappa la cui chiave corrisponde alle categorie Chiave dati.</p>
	<p>Torta su una mappa</p> <p>Calcola la somma di un campo continuo (Valori) in ciascuna categoria di un campo relativo alla categoria (Categorie) per ciascuna funzione della mappa (Chiave dati) e traccia una mappa e la somma come fette di un grafico a torta al centro di ciascuna funzione della mappa.</p> <p><i>Richiede:</i> una coppia di campi relativi alla categoria ed un campo continuo. Un file di mappa la cui chiave corrisponde alle categorie Chiave dati.</p>		<p>Grafico a linee su una mappa</p> <p>Calcola una statistica di riepilogo per un campo continuo (Y) per ciascun valore di un altro campo (X) per ciascuna funzione della mappa (chiave dati) e traccia una mappa e dei grafici a linee che collegano i valori al centro di ciascuna funzione della mappa.</p> <p><i>Richiede:</i> un campo relativo alla categoria ed un paio di campi di qualsiasi tipo. Un file di mappa la cui chiave corrisponde alle categorie Chiave dati.</p>
	<p>Coordinate su una mappa di riferimento</p> <p>Traccia una mappa e dei punti utilizzando i campi continui (Longitudine e Latitudine) che identificano le coordinate per i punti.</p> <p><i>Richiede:</i> una coppia di campi intervallo. Un file di mappa.</p>		<p>Freccie su una mappa di riferimento</p> <p>Traccia una mappa e le frecce utilizzando campi continui che identificano i punti di inizio (Inizio long e Inizio lat) ed i punti di fine (Fine long e Fine lat) per ciascuna freccia. Ciascun record/caso nei dati produce una freccia nella mappa.</p> <p><i>Richiede :</i> quattro campi continui. Un file di mappa.</p>

Tabella 45. Tipi di grafici disponibili (Continua).

Icona Grafico	Descrizione	Icona Grafico	Descrizione
	<p>Mappa di sovrapposizione punti</p> <p>Traccia una mappa di riferimento a cui sovrappone un'altra mappa di punti con funzioni di punti colorate da un campo relativo alla categoria (Colore).</p> <p><i>Richiede:</i> una coppia di campi relativi alla categoria. Un file di mappa a punti la cui chiave corrisponde alle categorie Chiave dati. Un file di mappa di riferimento.</p>		<p>Mappa di sovrapposizione dei poligoni</p> <p>Traccia una mappa di riferimento a cui sovrappone un'altra mappa dei poligoni con funzioni di poligono colorate da un campo relativo alla categoria (Colore).</p> <p><i>Richiede:</i> una coppia di campi relativi alla categoria. Un file di mappa dei poligoni la cui chiave corrisponde alle categorie Chiave dati. Un file di mappa di riferimento.</p>
	<p>Mappa di sovrapposizione linee</p> <p>Traccia una mappa di riferimento a cui sovrappone un'altra mappa di linee con funzioni di linea colorate da un campo relativo alla categoria (Colore).</p> <p><i>Richiede:</i> una coppia di campi relativi alla categoria. Un file di mappa di linea la cui chiave corrisponde alle categorie Chiave dati. Un file di mappa di riferimento.</p>		

Creazione di visualizzazioni delle mappe

Per molte visualizzazioni, è necessario effettuare solo due scelte: i campi (variabili) di interesse ed un modello per la visualizzazione di tali campi. Non sono necessarie ulteriori scelte né azioni. Le visualizzazioni della mappa richiedono almeno un passo aggiuntivo: la selezione di un file mappa che definisce le informazioni geografiche per la visualizzazione della mappa.

Di seguito sono riportati i passaggi di base per la creazione di una mappa semplice:

1. Selezionare i campi di interesse sulla scheda Di base. Per informazioni relative al tipo ed al numero di campi richiesti per le differenti visualizzazioni delle mappe, consultare "Tipi di visualizzazione della lavagna grafica incorporati disponibili" a pagina 176.
2. Selezionare un modello di mappa.
3. Fare clic sulla scheda Dettagliato.
4. Verificare che la **Chiave dati** e altri elenchi a discesa richiesti siano impostati sui campi corretti.
5. Nel gruppo File mappa, fare clic su **Seleziona un file mappa**.
6. Utilizzare la finestra di dialogo Seleziona mappe per scegliere il file mappa e la chiave mappa. I valori della chiave mappa devono corrispondere ai valori del campo specificato dalla **Chiave dati**. È possibile utilizzare il pulsante **Confronta** per confrontare questi valori. Se si seleziona un modello di mappa di sovrapposizione, si dovrà anche scegliere una mappa di riferimento. La mappa di

riferimento non è adattata ai dati. Viene utilizzata come sfondo per la mappa principale. Per ulteriori informazioni relative alla finestra di dialogo Seleziona mappe, consultare "Selezione dei file delle mappe per le visualizzazioni delle mappe" a pagina 174.

7. Fare clic su **OK** per chiudere la finestra di dialogo Seleziona mappe.
8. Nel Selezionatore modelli per lavagna grafica, fare clic su **Esegui** per creare la visualizzazione della mappa.

Lavagna grafica Esempi

In questa sezione vengono presentati numerosi esempi che illustrano le opzioni disponibili. Tali esempi forniscono inoltre informazioni per interpretare le visualizzazioni risultanti.

Viene utilizzato il flusso *graphboard.str*, che fa riferimento ai file di dati *employee_data.sav*, *customer_subset.sav* e *worldsales.sav*, disponibili nella cartella *Demos* di qualunque installazione di IBM SPSS Modeler Client. Tale cartella si trova nel gruppo di programma IBM SPSS Modeler del menu Start di Windows. Il file *graphboard.str* si trova nella cartella *streams*.

Si consiglia di leggere gli esempi nell'ordine proposto. Gli esempi successivi sono creati in base a quelli precedenti.

Esempio: Grafico a barre con statistica riassuntiva

Verrà generato un grafico a barre che riepiloga un campo numerico continuo/una variabile numerica continua per ciascuna categoria di un insieme/una variabile categoriale. Nel caso specifico, il grafico a barre mostrerà lo stipendio medio di uomini e donne.

Questo esempio e molti dei successivi utilizzano *Employee data*, un insieme di dati ipotetico contenente informazioni sui dipendenti di un'azienda.

1. Aggiungere un nodo origine File Statistics che punta a *employee_data.sav*.
2. Aggiungere un nodo Lavagna grafica e aprirlo per modificarlo.
3. Nella scheda Di base, selezionare *Sesso* e *Stipendio attuale*. Premere CTRL facendo clic con il pulsante sinistro del mouse per selezionare più campi/variabili.
4. Selezionare **A barre**.
5. Dall'elenco a discesa Riepilogo selezionare **Media**.
6. Fare clic su **Esegui**.
7. Nella visualizzazione risultante fare clic sul pulsante nella barra degli strumenti "Visualizza le etichette di valori e campi" (il secondo del gruppo di due al centro della barra degli strumenti).

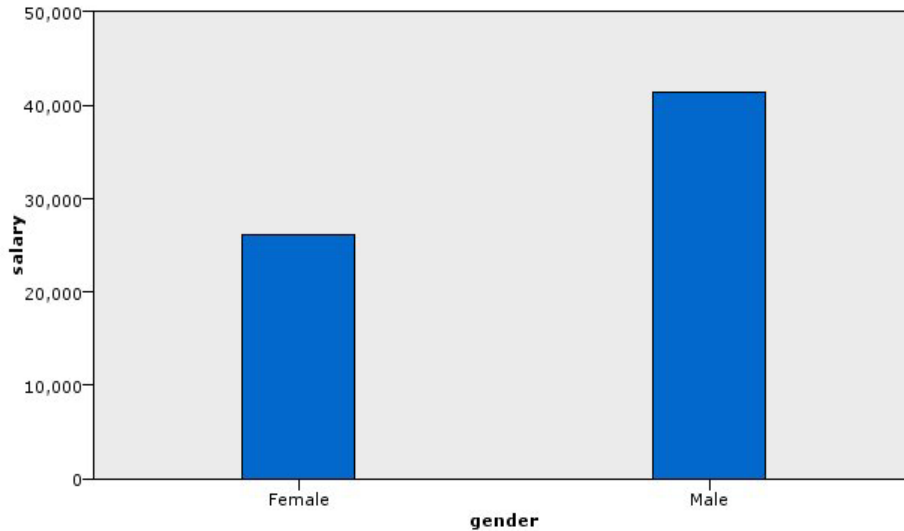


Figura 10. Grafico a barre con statistiche riassuntive

Si osserva quanto segue:

- Esaminando l'altezza delle barre, risulta evidente che lo stipendio medio degli uomini è maggiore di quello delle donne.

Esempio: Grafico a barre raggruppate con statistiche riassuntive

Si procederà ora alla creazione di un grafico a barre raggruppato per stabilire se la differenza tra lo stipendio medio degli uomini e quello delle donne dipende dal tipo di professione. È infatti possibile che, in media, le donne guadagnino di più rispetto agli uomini in determinati lavori.

Nota: questo esempio utilizza *Employee data*.

1. Aggiungere un nodo Lavagna grafica e aprirlo per modificarlo.
2. Nella scheda Di base, selezionare *Categoria lavorativa* e *Stipendio attuale*. Premere CTRL facendo clic con il pulsante sinistro del mouse per selezionare più campi/variabili.
3. Selezionare **A barre**.
4. Dall'elenco Riepilogo selezionare **Media**.
5. Fare clic sulla scheda Dettagliato. Come si vede, sono attive le selezioni effettuate nella scheda precedente.
6. Nel gruppo Elementi estetici facoltativi scegliere *Sesso* dall'elenco a discesa Colore.
7. Fare clic su **Esegui**.

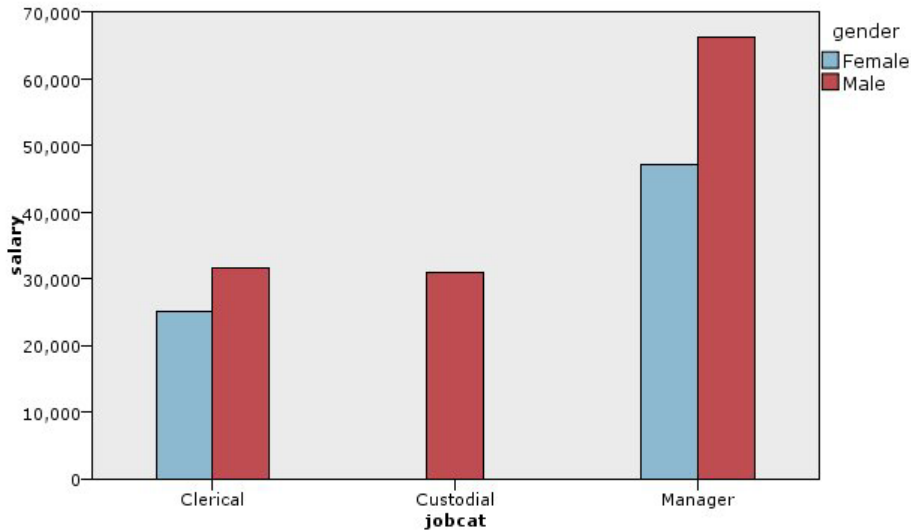


Figura 11. Grafico a barre raggruppato

Si osserva quanto segue:

- La differenza tra gli stipendi medi per ogni tipo di professione non è così significativa come nel grafico a barre che confrontava gli stipendi medi di tutti gli uomini e tutte le donne. Forse ciò è dovuto alla presenza di numeri variabili di uomini e donne in ciascun gruppo. Lo si può verificare generando un grafico a barre dei conteggi.
- Indipendentemente dal tipo di professione, lo stipendio medio degli uomini è sempre superiore a quello delle donne.

Esempio: Istogramma a riquadri

Verrà creato un istogramma suddiviso in riquadri in base al sesso, così da poter confrontare le distribuzioni di frequenza dello stipendio per gli uomini e per le donne. La distribuzione di frequenza indica il numero di casi/righe all'interno di specifici intervalli di stipendi. L'istogramma a riquadri può essere utile per approfondire l'analisi della differenza tra gli stipendi in base al sesso.

Nota: questo esempio utilizza *Employee data*.

1. Aggiungere un nodo Lavagna grafica e aprirlo per modificarlo.
2. Nella scheda Di base, selezionare *Stipendio attuale*.
3. Selezionare **Istogramma**.
4. Fare clic sulla scheda Dettagliato.
5. Nel gruppo Riquadro e animazione scegliere *sesso* dall'elenco a discesa Riquadro orizzontale.
6. Fare clic su **Esegui**.

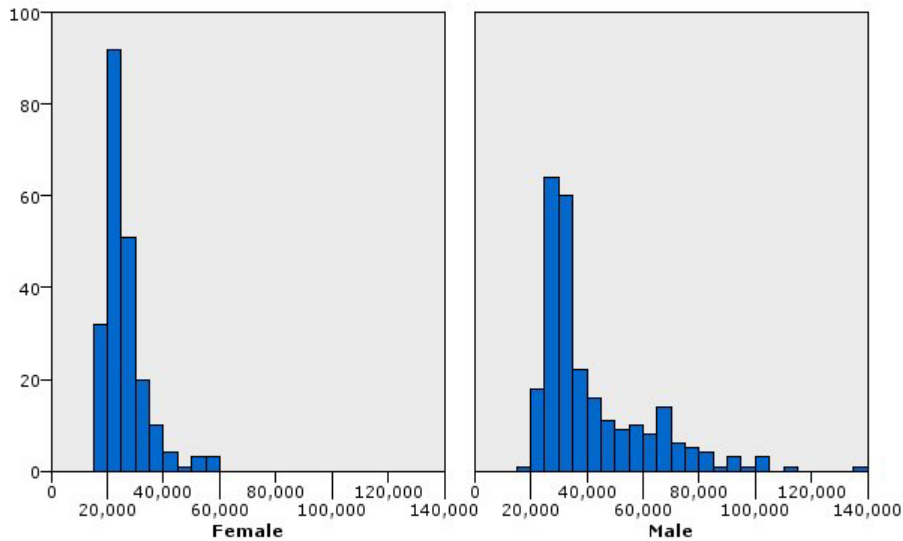


Figura 12. Istogramma a riquadri

Si osserva quanto segue:

- Nessuna delle distribuzioni di frequenza è normale. In altre parole, gli istogrammi non presentano curve a campana, come accadrebbe se i dati avessero una distribuzione normale.
- Le barre più lunghe sono sulla sinistra di ciascun gruppo. Pertanto, per entrambi i sessi, il numero di persone che guadagna stipendi più bassi è maggiore rispetto al numero di persone che può contare su salari più alti.
- Le distribuzioni di frequenza degli stipendi di uomini e donne non sono uguali. Si noti la forma degli istogrammi. Gli stipendi elevati sono più diffusi tra gli uomini che tra le donne.

Esempio: Grafico a punti con riquadri

Come gli istogrammi, i grafici a punti mostrano la distribuzione di un intervallo numerico continuo, con la differenza che mostrano tutte le righe/tutti i casi dei dati anziché visualizzare i conteggi relativi intervalli di dati in bin. Ecco perché un grafico a punti garantisce una maggiore granularità rispetto a un istogramma. Non a caso, quando si analizzano distribuzioni di frequenza è preferibile iniziare con un grafico a punti.

Nota: questo esempio utilizza *Employee data*.

1. Aggiungere un nodo Lavagna grafica e aprirlo per modificarlo.
2. Nella scheda Di base, selezionare *Stipendio attuale*.
3. Selezionare **Grafico a punti**.
4. Fare clic sulla scheda Dettagliato.
5. Nel gruppo Riquadro e animazione scegliere *Sesso* dall'elenco a discesa Riquadro orizzontale.
6. Fare clic su **Esegui**.
7. Ingrandire la finestra dell'output risultante per visualizzare il grafico in modo più chiaro.

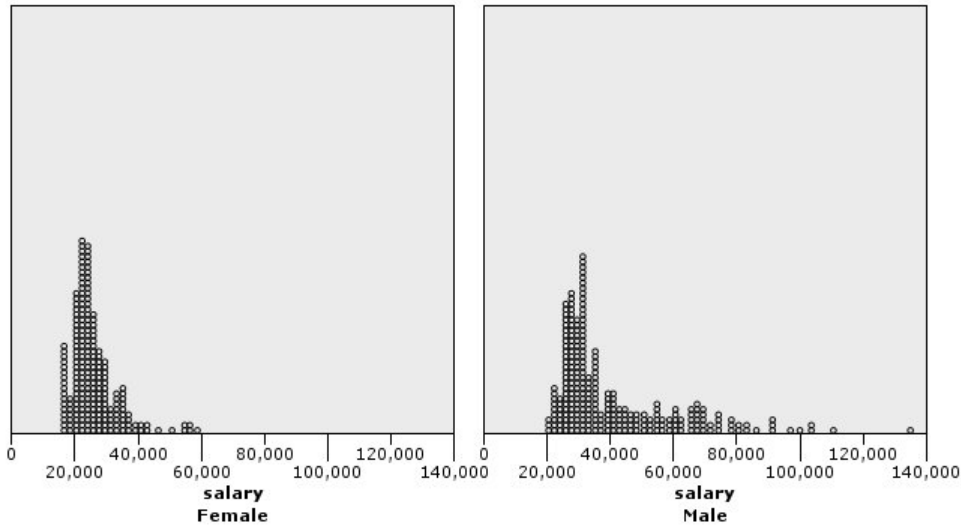


Figura 13. Grafico a punti con riquadri

Rispetto all'istogramma (consultare “Esempio: Iistogramma a riquadri” a pagina 186), è possibile osservare quanto riportato di seguito:

- Il picco (20.000) registrato nell'istogramma per le donne è meno evidente nel grafico a punti. Intorno a questo valore si concentrano molti casi/molte righe, ma gran parte di tali valori sono più vicini a 25.000. Questo livello di granularità non traspare nell'istogramma.
- Sebbene dall'istogramma relativo agli uomini si evinca che il loro stipendio medio tende a scendere gradualmente dopo i 40.000, il grafico a punti mostra che fino a 80.000 la distribuzione è abbastanza uniforme. In corrispondenza di ogni valore di stipendio compreso in tale intervallo, sono presenti tre o più uomini che guadagnano quel determinato salario.

Esempio: Grafico a scatole

I grafici a scatole rappresentano un altro utile metodo di visualizzazione della distribuzione dei dati. Un grafico di questo tipo contiene svariate misure statistiche che verranno discusse dopo aver preparato un esempio.

Nota: questo esempio utilizza *Employee data*.

1. Aggiungere un nodo Lavagna grafica e aprirlo per modificarlo.
2. Nella scheda Di base, selezionare *Sesso* e *Stipendio attuale* . Premere CTRL facendo clic con il pulsante sinistro del mouse per selezionare più campi/variabili.
3. Selezionare **Grafico a scatole**.
4. Fare clic su **Esegui**.

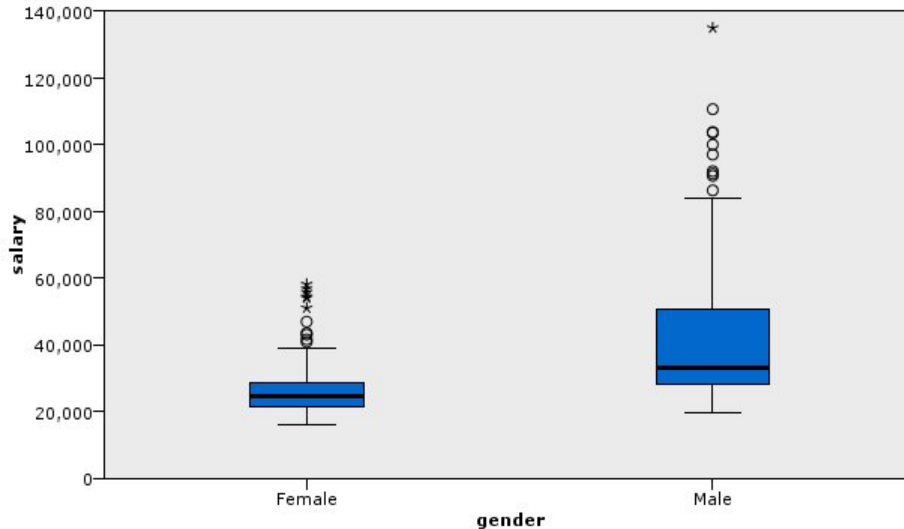


Figura 14. Grafico a scatole

È giunto il momento di esaminare le varie parti del grafico a scatole:

- La linea scura al centro delle scatole è la mediana dello *stipendio*. Il 50% dei casi/delle righe ha un valore maggiore rispetto alla mediana, l'altro 50% un valore minore. Così come la media, anche la mediana è una misura di tendenza centrale, con la differenza che è meno influenzata da casi/righe con valori estremi. In questo esempio, la mediana è inferiore alla media (consultare "Esempio: Grafico a barre con statistica riassuntiva" a pagina 184). La differenza tra media e mediana indica la presenza di alcuni casi/alcune righe con valori estremi che innalzano la media. In altri termini, c'è qualche dipendente con stipendi elevati.
- La parte inferiore della scatola indica il 25° percentile. Il 25% dei casi/delle righe mostra valori al di sotto del 25° percentile. La parte superiore della scatola rappresenta il 75° percentile. Il 25% dei casi/delle righe mostra valori al di sopra del 75° percentile. Ciò significa che il 50% dei casi/delle righe rientra all'interno della scatola. Quest'ultima risulta molto più corta per le donne, il che fa pensare che lo *stipendio* presenti meno variazioni per le donne rispetto a quello che avviene per gli uomini. Spesso la parte superiore e quella inferiore della scatola vengono denominate **cardini**.
- Le barre a T che si diramano dalle scatole sono dette **recinzioni interne** o **baffi**. Si estendono per 1,5 volte l'altezza della scatola oppure, se nell'intervallo non sono presenti casi/righe con valori assegnati, fino ai valori minimi o a quelli massimi. Se i dati hanno una distribuzione normale, si prevede che circa il 95% di essi sia compreso tra le recinzioni interne. In questo esempio, le recinzioni interne hanno un'estensione minore per le donne, un ulteriore indizio che lo *stipendio* varia meno per le donne che per gli uomini.
- I punti sono **valori anomali**, che cioè non rientrano nelle recinzioni interne. I valori anomali sono valori estremi. Gli asterischi o le stelle sono **valori anomali estremi** e rappresentano casi/righe con valori che superano di tre volte l'altezza delle scatole. Sono presenti numerosi valori anomali sia per le donne sia per gli uomini. È opportuno ricordare che la media è maggiore della mediana proprio a causa di questi valori anomali.

Esempio: Grafico a torta

A questo punto verrà utilizzato un altro insieme di dati per esaminare altri tipi di visualizzazione. Si tratta di *customer_subset*, un file di dati ipotetico contenente informazioni sui clienti.

Per prima cosa verrà generato un grafico a torta per verificare le percentuali di clienti in diverse regioni geografiche.

1. Aggiungere un nodo origine File Statistics che punta a *customer_subset.sav*.
2. Aggiungere un nodo Lavagna grafica e aprirlo per modificarlo.

3. Nella scheda Di base, selezionare *Indicatore geografico*.
4. Selezionare **Torta dei conteggi**.
5. Fare clic su **Esegui**.

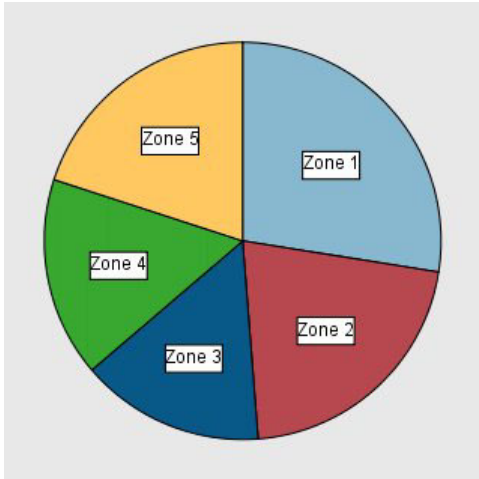


Figura 15. Grafico a torta

Si osserva quanto segue:

- La Zona 1 ha più clienti delle altre singole zone.
- I clienti sono equamente distribuiti nelle altre zone.

Esempio: Mappa termica

Verrà ora creata una mappa termica categoriale per controllare il reddito medio dei clienti in diverse aree geografiche e fasce di età.

Nota: in questo esempio, viene utilizzato *customer_subset*.

1. Aggiungere un nodo Lavagna grafica e aprirlo per modificarlo.
2. Nella scheda Di base, selezionare *Indicatore geografico*, *Categoria età* e *Reddito familiare in migliaia*, in questo ordine. Premere CTRL facendo clic con il pulsante sinistro del mouse per selezionare più campi/variabili.
3. Selezionare **Mappa termica**.
4. Fare clic su **Esegui**.
5. Nella finestra dell'output risultante, fare clic sul pulsante nella barra degli strumenti "Visualizza le etichette di valori e campi" (quello a destra dei due al centro della barra degli strumenti).

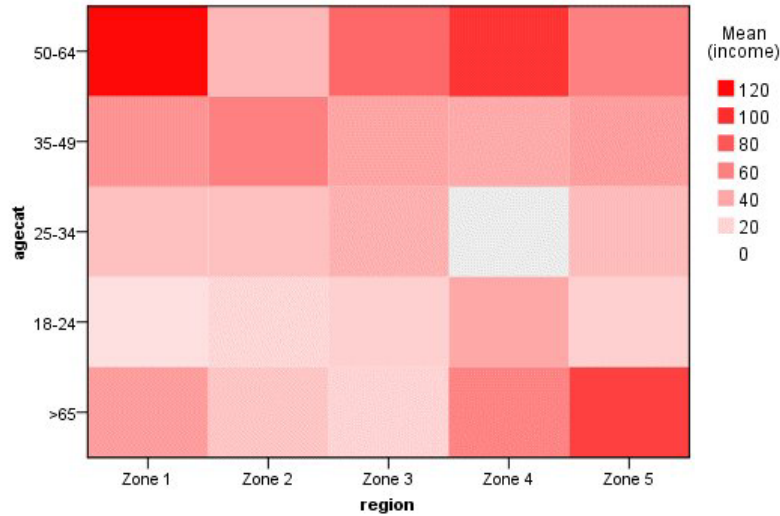


Figura 16. Mappa termica categoriale

Si osserva quanto segue:

- Le mappe termiche sono tabelle che usano colori anziché numeri per rappresentare i valori delle celle. Il rosso intenso indica il valore più elevato, il grigio un valore basso. Il valore di ciascuna cella è pari alla media del campo continuo/della variabile continua per ogni coppia di categorie.
- Ad eccezione della Zona 2 e della Zona 5, il gruppo di clienti con età compresa tra 50 e 64 hanno un reddito familiare medio maggiore rispetto agli altri gruppi.
- Non sono presenti clienti di età compresa tra 25 e 34 nella Zona 4.

Esempio: Matrice di grafici a dispersione (SPLOM)

Verrà creata una matrice di grafici a dispersione per molte variabili diverse, in modo da poter stabilire se esistono rapporti tra le variabili dell'insieme di dati.

Nota: in questo esempio, viene utilizzato *customer_subset*.

1. Aggiungere un nodo Lavagna grafica e aprirlo per modificarlo.
2. Nella scheda Di base, selezionare *Età in anni*, *Reddito familiare in entrata* e *Debito carta di credito in migliaia*. Premere CTRL facendo clic con il pulsante sinistro del mouse per selezionare più campi/variabili.
3. Selezionare **SPLOM**.
4. Fare clic su **Esegui**.
5. Ingrandire la finestra dell'output per visualizzare la matrice in modo più chiaro.

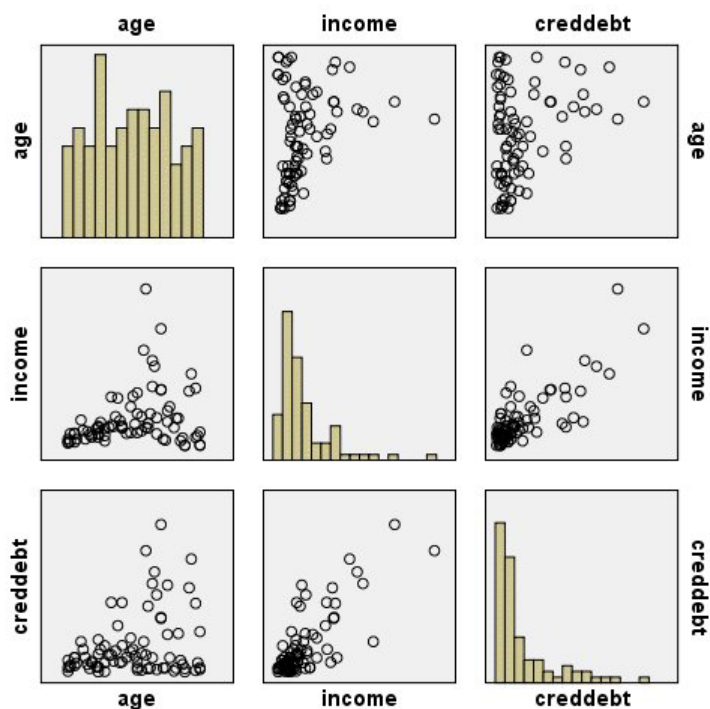


Figura 17. Matrice di grafici a dispersione (SPLOM)

Si osserva quanto segue:

- Gli istogrammi visualizzati sulla diagonale mostrano la distribuzione di ciascuna variabile nella SPLOM. L'istogramma relativo al valore *età* viene visualizzato nella cella in alto a sinistra, quello relativo ai *reddito* nella cella centrale e quello per il *creddeb* nella cella in basso a destra. Nessuna delle variabili ha una distribuzione normale. In altri termini, nessuno degli istogrammi presenta una curva a campana. Si noti inoltre che gli istogrammi per *reddito* e *creddeb* sono positivamente asimmetrici.
- Non sembra esserci alcuna relazione tra *età* e altre variabili.
- Esiste un rapporto lineare tra *reddito* e *creddeb*. Ovvero, *creddeb* aumenta con l'aumento del *reddito*. È possibile creare singoli grafici a dispersione per queste e per altre variabili correlate al fine di approfondire l'esame di questi rapporti.

Esempio: Coropleta (mappa colori) di somme

A questo punto verrà creata una visualizzazione mappa. Quindi, nell'esempio successivo, si procederà a creare una variazione di questa visualizzazione. L'insieme di dati è *worldsales*, un file di dati ipotetici contenente i ricavi derivanti dalle vendite suddivisi per continenti e prodotti.

1. Aggiungere un nodo Lavagna grafica e aprirlo per modificarlo.
2. Nella scheda Di base selezionare *Continente* e *Ricavo*. Premere CTRL facendo clic con il pulsante sinistro del mouse per selezionare più campi/variabili.
3. Selezionare **Coropleta di somme**.
4. Fare clic sulla scheda Dettagliato.
5. Nel gruppo Elementi estetici facoltativi scegliere *Continente* dall'elenco a discesa Etichetta dei dati.
6. Nel gruppo File mappa, fare clic su **Seleziona un file mappa**.
7. Nella finestra di dialogo Seleziona mappe, verificare che **Mappa** sia impostato su *Continenti* e **Chiave mappa** sia impostato su *CONTINENTE*.
8. Nei gruppi Confronta la mappa e i valori dei dati, fare clic su **Confronta** per verificare che le chiavi della mappa corrispondano alle chiavi dei dati. In questo esempio, tutti i valori delle chiavi dei dati hanno chiavi della mappa e caratteristiche corrispondenti. Inoltre, si nota che non vi sono dati relativi all'Oceania.

9. Nella finestra di dialogo Seleziona mappe, fare clic su **OK**.
10. Fare clic su **Esegui**.



Figura 18. Coropleta di somme

Da quella visualizzazione della mappa, si evince facilmente che i ricavi maggiori sono generati in America del Nord e i ricavi minori sono riferiti a Sud America e Africa. Ogni continente è etichettato in quanto è stato utilizzato *Continente* negli elementi estetici dell'etichetta dei dati.

Esempio: Grafici a barre su una mappa

Questo esempio mostra la suddivisione dei ricavi per prodotto nei singoli continenti.

Nota: questo esempio utilizza *worldsales*.

1. Aggiungere un nodo Lavagna grafica e aprirlo per modificarlo.
2. Nella scheda Di base, selezionare *Continente*, *Prodotto* e *Ricavo*. Premere CTRL facendo clic con il pulsante sinistro del mouse per selezionare più campi/variabili.
3. Selezionare **Barre su una mappa**.
4. Fare clic sulla scheda Dettagliato.
Quando si utilizza più di un campo di un tipo specifico, è importante verificare che ogni campo sia assegnato allo slot esatto.
5. Nell'elenco a discesa Categorie, selezionare *Prodotto*.
6. Nell'elenco a discesa Valori, selezionare *Ricavo*.
7. Nell'elenco a discesa Chiave dati, scegliere *Continente*.
8. Nell'elenco a discesa Riepilogo, selezionare *Somma*.
9. Nel gruppo File mappa, fare clic su **Seleziona un file mappa**.
10. Nella finestra di dialogo Seleziona mappe, verificare che **Mappa** sia impostato su *Continenti* e **Chiave mappa** sia impostato su *CONTINENTE*.

11. Nei gruppi Confronta la mappa e i valori dei dati, fare clic su **Confronta** per verificare che le chiavi della mappa corrispondano alle chiavi dei dati. In questo esempio, tutti i valori delle chiavi dei dati hanno chiavi della mappa e caratteristiche corrispondenti. Inoltre, si nota che non vi sono dati relativi all'Oceania.
12. Nella finestra di dialogo Seleziona mappe, fare clic su **OK**.
13. Fare clic su **Esegui**.
14. Ingrandire la finestra dell'output risultante per visualizzare l'immagine in modo più chiaro.

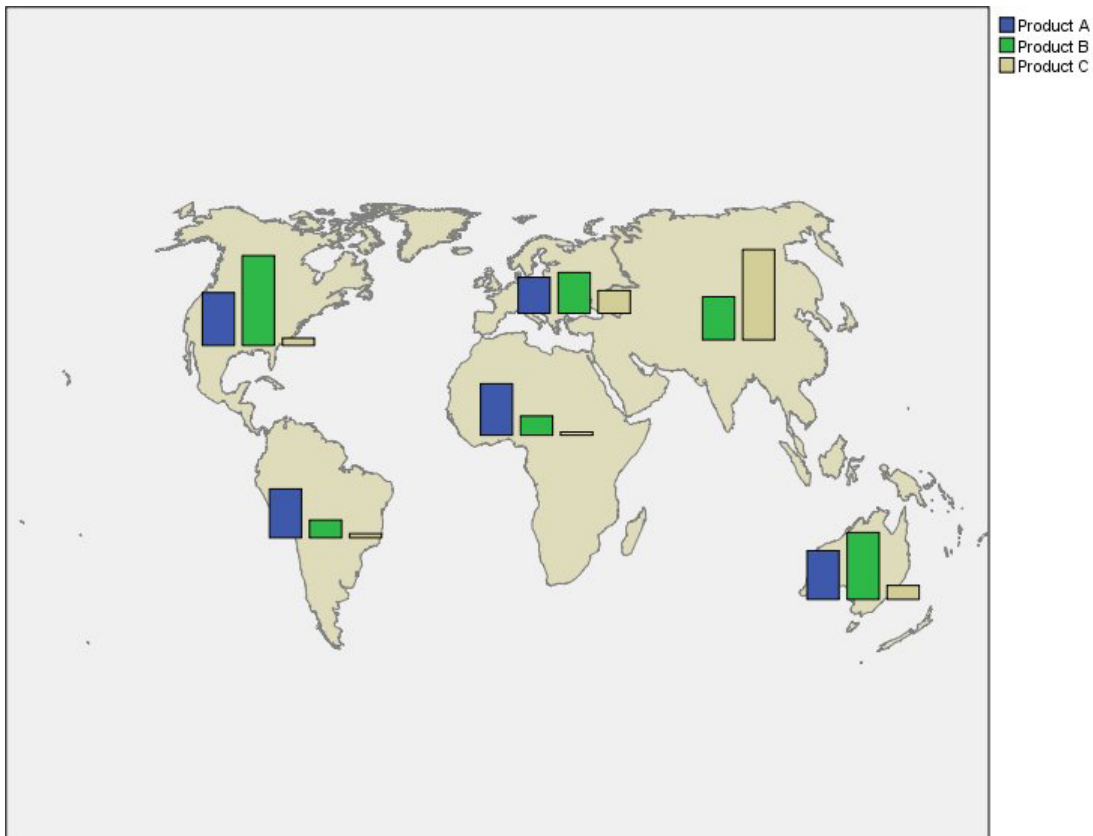


Figura 19. Grafici a barre su una mappa

Si osserva quanto segue:

- La distribuzione dei ricavi totali tra i prodotti è molto simile in Sud America e in Africa.
- Il *Prodotto C* genera il ricavo più basso ovunque tranne che in Asia.
- In Asia, il *Prodotto A* genera un ricavo minimo o null.

Scheda Aspetto del nodo Lavagna grafica

Prima di creare un grafico è possibile specificarne le opzioni dell'aspetto.

Opzioni generali dell'aspetto

Titolo. Immettere il testo da utilizzare come titolo del grafico.

Sottotitolo. Immettere il testo da utilizzare come sottotitolo del grafico.

Didascalia. Immettere il testo che verrà utilizzato per la didascalia del grafico.

Campionamento. Specificare un metodo per rappresentare insiemi di dati di notevoli dimensioni. È possibile specificare una dimensione massima per l'insieme di dati oppure utilizzare il numero di record di default. Per gli insiemi di dati di grandi dimensioni le prestazioni risultano migliori se si seleziona l'opzione **Campione**. In alternativa, è possibile rappresentare tutti i punti dati selezionando **Utilizza tutti i dati**, tuttavia occorre tenere presente che in questo modo le prestazioni del software potrebbero risultare notevolmente ridotte.

Opzioni dell'aspetto dei fogli di stile

Sono disponibili due pulsanti che consentono di controllare i modelli di visualizzazione (e fogli di stile e mappe) disponibili:

Gestisci. Gestire modelli, fogli di stile e mappe di visualizzazione sul computer. È possibile importare, esportare, rinominare ed eliminare modelli, fogli di stile e mappe di visualizzazione sul computer locale. Per ulteriori informazioni, consultare l'argomento "Gestione dei modelli, dei fogli di stile e dei file delle mappe" a pagina 196.

Posizione. Modificare la posizione in cui sono archiviati i modelli, i fogli di stile e le mappe di visualizzazione. La posizione corrente viene elencata alla destra del pulsante. Per ulteriori informazioni, consultare l'argomento "Impostazione della posizione di modelli, fogli di stile e mappe" a pagina 196.

L'esempio riportato di seguito mostra il punto in cui vengono posizionate le opzioni dell'aspetto in un grafico. *Nota:* non tutti i grafici utilizzano tutte le opzioni.

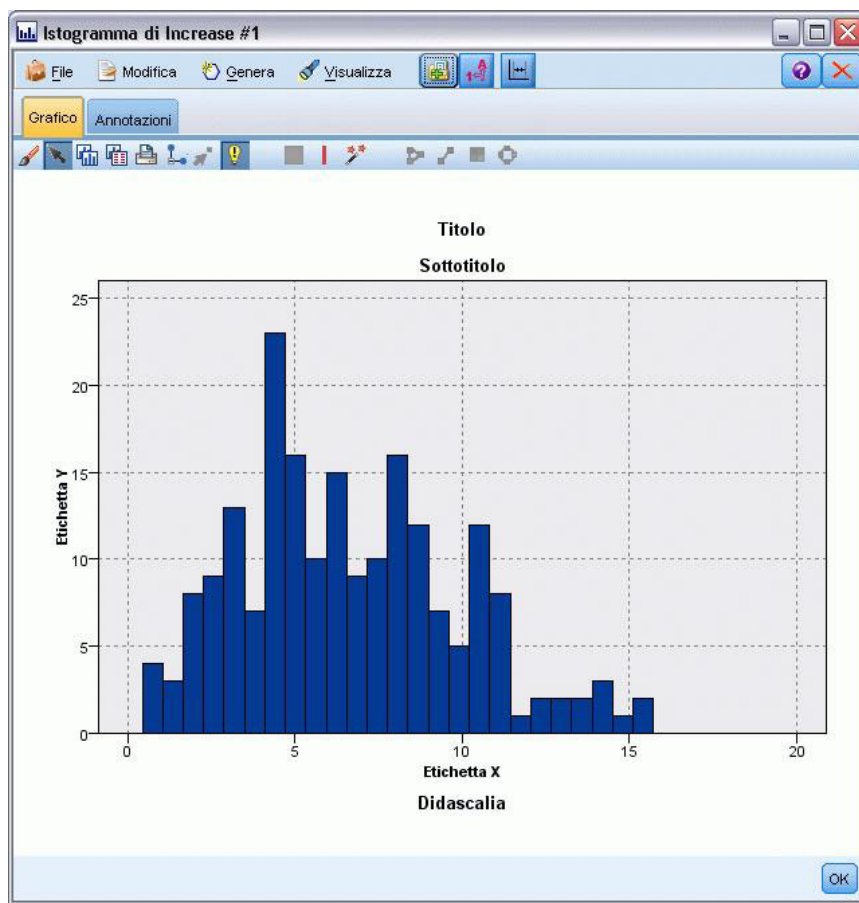


Figura 20. Posizione delle diverse opzioni dell'aspetto del grafico

Impostazione della posizione di modelli, fogli di stile e mappe

I modelli di visualizzazione, i fogli di stile di visualizzazione e i file delle mappe sono archiviati in una cartella locale specifica o nel IBM SPSS Collaboration and Deployment Services Repository. Quando si selezionano i modelli, i fogli di stile e le mappe, sono visualizzati solo quelli integrati in questo percorso. Il fatto di archiviare tutti i modelli, i fogli di stile e i file delle mappe nello stesso punto, consente alle applicazioni di IBM SPSS di potervi accedere più facilmente. Per informazioni sull'aggiunta di modelli, fogli di stile e file delle mappe in questa posizione, vedere "Gestione dei modelli, dei fogli di stile e dei file delle mappe".

Come impostare la posizione di modelli, fogli di stile e file delle mappe

1. Nella finestra di dialogo di un modello o foglio di stile, fare clic su **Posizione...** per visualizzare la finestra di dialogo Modelli, fogli di stile e mappe.
2. Selezionare un'opzione per impostare il percorso predefinito di modelli, fogli di stile e file delle mappe:

Computer locale. I modelli, i fogli di stile e i file delle mappe si trovano in una cartella specifica del computer locale. In Windows XP, si tratta della cartella `C:\Documents and Settings\<utente>\Application Data\SPSSInc\Graphboard`. Non è possibile modificare questa cartella.

IBM SPSS Collaboration and Deployment Services Repository. I modelli, i fogli di stile ed i file mappa sono contenuti in una cartella specificata dall'utente in IBM SPSS Collaboration and Deployment Services Repository. Per individuare la cartella specifica, fare clic su **Cartella**. Per ulteriori informazioni, consultare "Utilizzo di IBM SPSS Collaboration and Deployment Services Repository come posizione per modelli, fogli di stile e file delle mappe".

3. Fare clic su **OK**.

Utilizzo di IBM SPSS Collaboration and Deployment Services Repository come posizione per modelli, fogli di stile e file delle mappe

I modelli e fogli di stile di visualizzazione possono essere memorizzati in IBM SPSS Collaboration and Deployment Services Repository. Si tratta di una cartella specifica in IBM SPSS Collaboration and Deployment Services Repository. Se tale percorso è impostato come posizione di default, è possibile selezionare tutti i modelli, i fogli di stile e i file delle mappe presenti al suo interno.

Come impostare una cartella in IBM SPSS Collaboration and Deployment Services Repository come percorso per i modelli, i fogli di stile e i file delle mappe

1. Nelle finestre di dialogo con il pulsante Posizione, fare clic su **Posizione....**
2. Selezionare IBM SPSS Collaboration and Deployment Services Repository.
3. Fare clic su **Cartella**.

Nota: se non si è già connessi a IBM SPSS Collaboration and Deployment Services Repository, verranno richieste le informazioni relative alla connessione.

4. Nella finestra di dialogo Seleziona cartella, selezionare la cartella in cui sono memorizzati i modelli, i fogli di stile e i file delle mappe.
5. Come opzione, è possibile selezionare un'etichetta da **Recupera etichetta**. Verranno visualizzati solo i modelli, i fogli di stile e i file delle mappe con l'etichetta selezionata.
6. Se si cerca una cartella contenente un determinato modello, foglio di stile o file delle mappe, è possibile effettuare una ricerca di questi elementi nella scheda Cerca. La finestra di dialogo Seleziona cartella seleziona automaticamente la cartella in cui sono contenuti il modello, il foglio di stile o il file di mappa.
7. Fare clic su **Seleziona cartella**.

Gestione dei modelli, dei fogli di stile e dei file delle mappe

È possibile gestire i modelli, i fogli di stile ed i file mappa in locale sul proprio computer utilizzando la finestra di dialogo Gestisci modelli, fogli di stile e mappe. Questa finestra di dialogo consente di

importare, esportare, rinominare ed eliminare i modelli, i fogli di stile di visualizzazione e i file delle mappe che si trovano nella posizione locale del computer.

Fare clic su **Gestisci...** in una delle finestre di dialogo in cui si selezionano i modelli, i fogli di stile o le mappe.

Finestra di dialogo Gestisci modelli, fogli di stile e mappe

Nella scheda Modello sono elencati tutti i modelli locali. Nella scheda Foglio di stile sono elencati tutti i fogli di stile locali e le visualizzazioni di esempio contenenti dei dati campione. È possibile selezionare uno dei fogli di stile e applicare i relativi stili alle visualizzazioni di esempio. Per ulteriori informazioni, consultare l'argomento "Applicazione di fogli di stile" a pagina 267. La scheda Mappa elenca tutti i file di mappa locali. Questa scheda visualizza anche le chiavi delle mappe compresi valori campione, un commento se ne è stato specificato uno durante la creazione della mappa e un'anteprima della mappa.

I pulsanti seguenti funzionano in qualsiasi scheda attiva in un determinato momento.

Importa. Consente di importare un modello, un foglio di stile di visualizzazione o un file mappa dal file system. Dopo l'importazione, un modello, un foglio di stile o un file delle mappe sono disponibili nell'applicazione IBM SPSS. Se si riceve un modello, un foglio di stile o un file delle mappe da un altro utente, è necessario importarlo prima di poterlo usare nella propria applicazione.

Esporta. Consente di esportare un modello, un foglio di stile di visualizzazione o un file mappa nel file system. Se si desidera inviare un modello, foglio di stile o un file delle mappe a un altro utente, è necessario esportarlo.

Rinomina. Consente di modificare il nome di un modello, foglio di stile di visualizzazione o file delle mappe. Non è consentito usare un nome già esistente.

Esporta chiave mappa. Consente di esportare le chiavi mappa in un file di valori separati da virgola (CSV). Questo pulsante è abilitato solo nella scheda Mappa.

Elimina. Consente di eliminare i modelli, i fogli di stile di visualizzazione o i file delle mappe selezionati. È possibile selezionare più modelli, fogli di stile o file delle mappe con Ctrl-clic. L'eliminazione non può essere annullata, pertanto si consiglia di agire con cautela.

Conversione e distribuzione degli shapefile delle mappe

Il Selezionatore modelli per lavagna grafica consente di creare visualizzazioni delle mappe a partire dalla combinazione di un modello di visualizzazione e di un file SMZ. I file SMZ sono simili agli shapefile (formato file SHP) ESRI in quanto entrambi contengono le informazioni geografiche per disegnare una mappa (ad esempio, i confini di stato), ma i primi sono ottimizzati per la visualizzazione delle mappe. Il Selezionatore modelli per lavagna grafica è preinstallato con una serie di file SMZ. Se per le visualizzazioni delle mappe si desidera utilizzare uno shapefile ESRI di cui si dispone già, è necessario convertire questo shapefile in un file SMZ utilizzando il programma di utilità di conversione mappe. Il programma di utilità di conversione mappe supporta gli shapefile ESRI di punti, polilinee o poligoni (tipi di forme 1, 3 e 5) contenenti un singolo strato.

Oltre alla conversione di shapefile ESRI, il programma di utilità di conversione mappe consente di modificare il livello di dettaglio della mappa, cambiare le etichette delle caratteristiche, unire e spostare le caratteristiche, e molte altre opzioni facoltative. È inoltre possibile utilizzare il programma di utilità di conversione mappe per modificare un file SMZ esistente (inclusi i file preinstallati).

Modifica di file SMZ preinstallati

1. Esportare il file SMZ dal sistema di gestione. Per ulteriori informazioni, consultare l'argomento "Gestione dei modelli, dei fogli di stile e dei file delle mappe" a pagina 196.

2. Utilizzare il programma di utilità di conversione mappe per aprire e modificare il file SMZ esportato. Si consiglia di salvare il file con un nome diverso. Per ulteriori informazioni, consultare l'argomento "Utilizzo del programma di utilità di conversione mappe" a pagina 199.
3. Importare il file SMZ modificato nel sistema di gestione. Per ulteriori informazioni, consultare l'argomento "Gestione dei modelli, dei fogli di stile e dei file delle mappe" a pagina 196.

Ulteriori risorse per file delle mappe

Dati geospaziali in formato di file SHP, che possono essere utilizzati per soddisfare requisiti riferiti alle mappe, sono disponibili presso molte fonti di dati pubbliche e private. Controllare i siti Web delle autorità locali se si cercano dati gratuiti. Molti dei modelli di questo prodotto sono basati su dati pubblicamente disponibili ottenuti da GeoCommons () e dallo U.S. Census Bureau (<http://www.census.gov>).

AVVISO IMPORTANTE: le informazioni relative a prodotti non IBM sono state ottenute dai fornitori di tali prodotti, da loro annunci pubblicati o da altre fonti disponibili al pubblico. IBM non ha verificato tali prodotti e non può confermare l'accuratezza delle prestazioni, la compatibilità o qualsiasi altra dichiarazione relativa a prodotti non IBM. Eventuali domande in merito alle funzionalità dei prodotti non IBM vanno indirizzate ai fornitori di tali prodotti. Qualsiasi riferimento nelle presenti informazioni a siti Web non IBM viene fornito esclusivamente per facilitare la consultazione e non rappresenta in alcun modo un'approvazione o sostegno da parte nostra di tali siti Web. Il materiale fornito in tali siti Web non fa parte del materiale relativo a questo programma IBM, salvo quando espressamente indicato in un file di Note legali che accompagna questo programma IBM, e l'utente utilizzerà il materiale di questi siti a proprio rischio e pericolo.

Concetti principali delle mappe

Per utilizzare in modo efficace il programma di utilità di conversione mappe è importante capire alcuni concetti chiave relativi agli shapefile.

Uno **shapefile** ingloba le informazioni geografiche per disegnare una mappa. Sono tre i tipi di shapefile supportati dal programma di utilità di conversione mappe:

- **Di punti.** Lo shapefile identifica le posizioni dei punti, ad esempio le città.
- **Di polilinee.** Lo shapefile identifica i percorsi e le relative posizioni, ad esempio i fiumi.
- **Di poligoni.** Lo shapefile identifica le aree delimitate e le relative posizioni, ad esempio gli stati.

Nella maggior parte dei casi si utilizza lo shapefile di poligoni. Le coroplete sono create a partire da shapefile di poligoni e utilizzano il colore per rappresentare un valore all'interno dei singoli poligoni (regioni). Gli shapefile di punti e di polilinee sono solitamente sovrapposti a uno shapefile di poligoni. Un esempio tipico è uno shapefile di punti delle città statunitensi sovrapposto a uno shapefile di poligoni degli stati degli USA.

Uno shapefile è composto da **caratteristiche**. Le caratteristiche sono le singole entità geografiche ad esempio, gli stati, le province, le città e così via. Lo shapefile contiene inoltre i dati relativi alle caratteristiche. Tali dati sono archiviati negli **attributi**. Gli attributi sono simili ai campi o alle variabili di un file di dati. Per una caratteristica esiste almeno un attributo che rappresenta la **chiave mappa** della caratteristica. La chiave mappa può essere un'etichetta, ad esempio un nome di stato o di provincia. La chiave mappa sarà l'elemento da collegare a una variabile/un campo in un file di dati per creare una visualizzazione mappa.

Si noti che sarà possibile conservare solo l'attributo o gli attributi chiave nel file SMZ. Il programma di utilità di conversione mappe, infatti, non supporta il salvataggio di altri attributi. Ciò significa che sarà necessario creare più file SMZ se si desidera eseguire un'aggregazione a diversi livelli. Ad esempio, se si desidera aggregare gli stati e le regioni degli Stati Uniti, sono necessari file SMZ separati: uno con una chiave che identifica gli stati e l'altro con una chiave che identifica le regioni.

Utilizzo del programma di utilità di conversione mappe

Come avviare il programma di utilità di conversione mappe

Dai menu, scegliere:

Strumenti > Programma di utilità di conversione mappe

Il programma di utilità di conversione mappe consta di quattro schermate principali (fasi). Una di queste fasi comprende delle fasi secondarie che consentono di modificare ulteriormente il file mappa.

Fase 1 - Scegliere il file di destinazione e il file di origine

Innanzitutto, è necessario selezionare un file mappa sorgente e una destinazione per il file mappa convertito. Per lo shapefile è necessario avere sia il file *.shp* sia il file *.dbf*.

Selezionare un file .shp (ESRI) o .smz da convertire. Individuare un file mappa esistente sul proprio computer. Si tratta del file che verrà convertito e salvato come file SMZ. Il file *.dbf* per lo shapefile *deve* essere archiviato nella stessa posizione con un nome file di base che corrisponda al file *.shp*. Il file *.dbf* è necessario perché contiene le informazioni sugli attributi per il file *.shp*.

Impostare una destinazione e un nome file per il file mappa convertito. Specificare un percorso e un nome file per il file SMZ che verrà creato a partire dalla sorgente mappa originale.

- **Importare nel selezionatore modelli.** Oltre a salvare un file nel file system, è possibile aggiungere la mappa all'elenco Gestisci del Selezionatore modelli. Se si sceglie questa opzione, la mappa sarà automaticamente disponibile nel Selezionatore modelli per i prodotti IBM SPSS installati sul proprio computer. Se l'importazione nel selezionatore modelli non viene eseguita ora, sarà necessario eseguire l'importazione manualmente in un secondo momento. Per ulteriori informazioni relative all'importazione delle mappe nel sistema di gestione del selezionatore modelli, consultare "Gestione dei modelli, dei fogli di stile e dei file delle mappe" a pagina 196.

Fase 2 - Scegliere la chiave mappa

Ora è necessario scegliere quali chiavi mappa includere nel file SMZ. In un secondo tempo sarà possibile modificare alcune opzioni che incideranno sull'interpretazione della mappa. Le fasi successive nel programma di utilità di conversione mappe includono un'anteprima della mappa. Le opzioni di interpretazione scelte verranno utilizzate per generare l'anteprima della mappa.

Scegli la chiave mappa principale. Selezionare l'attributo che rappresenta la chiave primaria per identificare ed etichettare le caratteristiche nella mappa. Ad esempio, la chiave primaria di una mappa della terra potrebbe essere l'attributo che identifica i nomi degli stati. La chiave primaria, inoltre, collegherà i propri dati alle caratteristiche della mappa, pertanto assicurarsi che i valori (etichette) dell'attributo che si sceglie corrisponderanno ai valori nei propri dati. Quando si sceglie un attributo vengono visualizzate alcune etichette di esempio. Se necessario, sarà possibile modificare queste etichette in una fase successiva.

Scegli le chiavi aggiuntive da includere. Oltre alla chiave mappa principale, selezionare tutti gli altri attributi chiave che si desidera includere nel file SMZ generato. Ad esempio, alcuni attributi potrebbero contenere etichette tradotte. Se si prevedono dati codificati in altre lingue, è utile conservare questi attributi. Si noti che è possibile scegliere solo quelle chiavi aggiuntive che rappresentano le stesse caratteristiche della chiave primaria. Per fare un esempio: se la chiave primaria è rappresentata dai nomi completi degli stati USA, è possibile selezionare solo le chiavi alternative che rappresentano gli stati USA, ad esempio le abbreviazioni.

Livella automaticamente la mappa. Gli shapefile con poligoni contengono solitamente troppi punti dati e troppi dettagli per le visualizzazioni delle mappe statistiche. L'eccesso di dettagli può creare confusione e incidere negativamente sulle prestazioni. È possibile ridurre il livello di dettaglio e generalizzare la mappa attraverso il livellamento. Dopo il livellamento la mappa risulterà più uniforme e l'interpretazione

sarà più veloce. Quando la mappa viene livellata automaticamente, l'angolo massimo è 15 gradi e la percentuale da conservare è 99. Per informazioni relative a tali impostazioni, consultare "Livella la mappa". Si noti che è possibile applicare un ulteriore livellamento in una fase successiva.

Elimina i confini tra i poligoni a contatto nella stessa caratteristica. Alcune caratteristiche possono includere delle sottocaratteristiche che hanno confini interni rispetto alle caratteristiche principali di interesse. Ad esempio, una mappa dei continenti potrebbe contenere confini interni relativi agli stati che compongono ciascun continente. Se si sceglie questa opzione, i confini interni non appariranno sulla mappa. Nell'esempio della mappa dei continenti, la scelta di questa opzione rimuove i confini statali, ma mantiene i confini continentali.

Fase 3 - Modificare la mappa

Ora che sono state specificate le opzioni di base della mappa, è possibile modificare alcune opzioni più specifiche. Queste modifiche sono facoltative. La presente fase del programma di utilità di conversione mappe illustra le attività associate e visualizza un'anteprima della mappa in modo da poter verificare le modifiche apportate. Alcune attività potrebbero non essere disponibili a seconda del tipo di shapefile (di punti, di polilinee o di poligoni) e del sistema di coordinate.

Ogni attività presenta i seguenti comandi comuni sul lato sinistro del programma di utilità di conversione mappe.

Mostra le etichette sulla mappa. Per impostazione predefinita, le etichette delle caratteristiche non vengono visualizzate nell'anteprima. È possibile scegliere di visualizzarle. Se da un lato le etichette sono utili nell'identificare le caratteristiche, dall'altro potrebbero interferire con la selezione diretta sulla mappa di anteprima. Selezionare questa opzione solo quando necessario, ad esempio quando si stanno modificando le etichette delle caratteristiche.

Colora l'anteprima della mappa. Per impostazione predefinita, l'anteprima della mappa visualizza le aree con un colore a tinta unita. Viene utilizzato lo stesso colore per tutte le caratteristiche. È possibile scegliere di avere un assortimento di colori assegnato alle singole caratteristiche della mappa. Questa opzione può essere utile nel distinguere le diverse caratteristiche nella mappa, ma soprattutto quando si uniscono le caratteristiche e si desidera vedere come le nuove caratteristiche sono rappresentate nell'anteprima.

Inoltre ogni attività presenta il seguente comando comune sul lato destro del programma di utilità di conversione mappe.

Annulla. Fare clic su **Annulla** per tornare allo stato precedente. È possibile annullare un massimo di 100 modifiche.

Livella la mappa: Gli shapefile con poligoni contengono solitamente troppi punti dati e troppi dettagli per le visualizzazioni delle mappe statistiche. L'eccesso di dettagli può creare confusione e incidere negativamente sulle prestazioni. È possibile ridurre il livello di dettaglio e generalizzare la mappa attraverso il livellamento. Dopo il livellamento la mappa risulterà più uniforme e l'interpretazione sarà più veloce. Questa opzione non è disponibile per le mappe di punti e di polilinee.

Max. angolo. L'angolo massimo, che deve avere un valore compreso tra 1 e 20, specifica la tolleranza per il livellamento di insiemi di punti che sono quasi lineari. Un valore più alto implica una maggiore tolleranza per il livellamento lineare e pertanto eliminerà più punti, consentendo così di ottenere una mappa più generalizzata. Per applicare il livellamento lineare, il programma di utilità di conversione mappe controlla l'angolo interno formato da ogni insieme di tre punti nella mappa. Se da 180 si sottrae l'angolo e il risultato è inferiore al valore specificato, allora il programma di utilità di conversione mappe eliminerà il punto intermedio. In altre parole, il programma di utilità di conversione mappe sta controllando se la linea formata dai tre punti è quasi una retta. Se tale linea è quasi una retta, il programma di utilità di conversione mappe la tratta come una linea retta tra i punti finali ed elimina il punto intermedio.

Percentuale da conservare. La percentuale da conservare, che deve avere un valore compreso tra 90 e 100, determina la quantità di area da mantenere quando la mappa viene livellata. Questa opzione incide solo su quelle caratteristiche che hanno più poligoni, come ad esempio nel caso di una caratteristica che include delle isole. Se dall'area totale di una caratteristica si sottrae un poligono e il risultato è maggiore della percentuale specificata dell'area originale, il programma di utilità di conversione mappe eliminerà il poligono dalla mappa. Il programma di utilità di conversione mappe non rimuoverà mai tutti i poligoni per la caratteristica. In altre parole, ci sarà sempre almeno un poligono per la caratteristica, a prescindere dalla quantità di livellamento applicata.

Dopo aver scelto un angolo massimo e una percentuale da conservare, fare clic su **Applica**. L'anteprima aggiornerà le modifiche del livellamento. Se è necessario livellare nuovamente la mappa, ripetere l'operazione fino a raggiungere il grado di livellamento desiderato. Si noti che esiste un limite al numero di livellamenti effettuabili. Dopo un certo numero di livellamenti, si raggiungerà un punto in cui non sarà possibile applicare ulteriori livellamenti alla mappa.

Modifica le etichette delle caratteristiche: È possibile modificare le etichette delle caratteristiche in base alle necessità, ad esempio per far corrispondere i propri dati, nonché riposizionare le etichette nella mappa. Anche se si pensa di non dover modificare le etichette, si consiglia comunque di rivederle prima di creare le visualizzazioni dalla mappa. Inoltre, poiché per impostazione predefinita l'anteprima non mostra le etichette, potrebbe essere utile selezionare **Mostra le etichette sulla mappa** per visualizzarle.

Chiavi. Selezionare la chiave contenente le etichette delle caratteristiche che si desidera rivedere e/o modificare.

Caratteristiche. Questo elenco visualizza le etichette delle caratteristiche contenute nella chiave selezionata. Fare doppio clic sull'etichetta nell'elenco per modificarla. Se le etichette sono visualizzate sulla mappa, è anche possibile fare doppio clic sulle etichette delle caratteristiche direttamente nell'anteprima della mappa. Nel caso si desideri confrontare le etichette rispetto a un file di dati effettivo, fare clic su **Confronta**.

X/Y. Questa caselle di testo elencano il punto centrale corrente dell'etichetta della funzione selezionata nella mappa. Le unità sono visualizzate nelle coordinate della mappa. Può trattarsi di coordinate cartesiane locali (ad esempio lo State Plane Coordinate System americano) o di coordinate geografiche (dove la X è la longitudine e la Y la latitudine). Specificare le coordinate per la nuova posizione dell'etichetta. Se le etichette sono visualizzate, è anche possibile fare clic su un'etichetta della mappa e trascinarla per spostarla. Le caselle di testo verranno aggiornate con la nuova posizione.

Confronta. Se si ha un file di dati contenente i valori dei dati che devono corrispondere alle etichette delle caratteristiche per una specifica chiave, fare clic su **Confronta** per visualizzare la finestra di dialogo Confronta con una sorgente dati esterna. In questa finestra di dialogo sarà possibile aprire il file di dati e confrontarne i valori direttamente con quelli nelle etichette delle caratteristiche della chiave mappa.

Finestra di dialogo Confronta con una sorgente dati esterna: La finestra di dialogo Confronta con una origine dati esterna consente di aprire un file di valori separati da tabulazione (con estensione *.txt*), un file di valori separati da virgole (con estensione *.csv*) o un di dati formattato per IBM SPSS Statistics (con estensione *.sav*). Quando il file è aperto, è possibile selezionare un campo nel file di dati da confrontare con le etichette delle caratteristiche in una chiave mappa specifica. Successivamente è possibile correggere eventuali discrepanze nel file mappa.

Campi nel file di dati. Scegliere il campo di cui si desidera confrontare i valori con le etichette delle caratteristiche. Se la prima riga nel file *.txt* o *.csv* contiene etichette descrittive per ciascun campo, selezionare l'opzione **Utilizza prima riga come etichette di colonna**. In caso contrario, ciascun campo verrà identificato dalla sua posizione nel file di dati (ad esempio, "Colonna 1", "Colonna 2" e così via).

Chiave da confrontare. Scegliere la chiave mappa di cui si desidera confrontare le etichette delle caratteristiche con i valori dei campi del file di dati.

Confronta. Fare clic su questo pulsante quando si è pronti a confrontare i valori.

Risultati del confronto. Per impostazione predefinita, la tabella Risultati del confronto elenca solo i valori dei campi che non hanno corrispondenza nel file di dati. L'applicazione prova a trovare un'etichetta di caratteristica correlata, solitamente verificando che non siano stati inseriti o rimossi degli spazi. Fare clic sull'elenco a discesa nella colonna *Etichetta mappa* per far corrispondere l'etichetta della caratteristica nel file mappa con il valore del campo visualizzato. Se non sono presenti etichette della caratteristica corrispondenti nel file mappa, scegliere *Lascia senza corrispondenza*. Se si desidera visualizzare i valori di tutti i campi, anche quelli che già corrispondono ad un'etichetta di caratteristiche, annullare la selezione di **Visualizza solo i casi senza corrispondenza**. È possibile eseguire questa operazione se si desidera sovrascrivere una o più corrispondenze.

Ciascuna caratteristica può essere utilizzata una sola volta per stabilire una corrispondenza con il valore di un campo. Se si desidera far corrispondere più caratteristiche a un unico valore di campo, è possibile unire le caratteristiche e successivamente far corrispondere la nuova caratteristica unita al valore del campo. Per ulteriori informazioni relative all'unione delle funzioni, consultare "Unisci caratteristiche".

Unisci caratteristiche: L'unione delle caratteristiche è utile per creare regioni più ampie in una mappa. Ad esempio, se si sta convertendo una mappa di stati, è possibile unire gli stati (le caratteristiche in questo esempio) in regioni più ampie, quali Nord, Sud, Est e Ovest.

Chiavi. Selezionare la chiave mappa contenente le etichette delle caratteristiche che consentiranno di identificare le caratteristiche da unire.

Caratteristiche. Fare clic sulla prima caratteristica che si desidera unire. Premere Ctrl e contemporaneamente fare clic sulle altre caratteristiche da unire. Si noti che le caratteristiche saranno selezionate anche nell'anteprima della mappa. È possibile fare clic e usare la combinazione Ctrl + clic sulle caratteristiche direttamente nell'anteprima della mappa, oltre a poterle selezionare dall'elenco.

Dopo aver selezionato le caratteristiche da unire, fare clic su **Unisci** per visualizzare la finestra di dialogo Assegnare un nome alla caratteristica creata con l'unione, da dove è possibile applicare un'etichetta alla nuova caratteristica. Per assicurarsi che i risultati soddisfino le attese, è possibile selezionare **Colora l'anteprima della mappa** dopo aver unito le caratteristiche.

Dopo l'unione delle caratteristiche si consiglia anche di spostare l'etichetta della nuova caratteristica. È possibile farlo nell'attività *Modifica le etichette delle caratteristiche*. Per ulteriori informazioni, consultare l'argomento "Modifica le etichette delle caratteristiche" a pagina 201.

Finestra di dialogo Assegnare un nome alla caratteristica creata con l'unione: La finestra di dialogo Assegnare un nome alla caratteristica creata con l'unione consente di assegnare le etichette alla nuova caratteristica unita.

La tabella Etichette visualizza le informazioni di ciascuna chiave nel file mappa e consente di assegnare un'etichetta per ciascuna chiave.

Nuova etichetta. Immettere una nuova etichetta per la caratteristica unita da assegnare alla chiave mappa specifica.

Chiave. La chiave mappa alla quale si sta assegnando la nuova etichetta.

Vecchie etichette. Le etichette delle caratteristiche che verranno unite nella nuova caratteristica.

Elimina confini tra poligoni a contatto. Selezionare questa opzione per rimuovere i confini dalle caratteristiche che sono state unite. Ad esempio, se si sono uniti gli stati in regioni geografiche più ampie, questa opzione rimuoverà i confini dei singoli stati.

Sposta caratteristiche: È possibile spostare le caratteristiche nella mappa. Può essere utile spostarle quando si desidera riunire delle caratteristiche, ad esempio la terra ferma e le relative isole.

Chiavi. Selezionare la chiave mappa contenente le etichette delle caratteristiche che consentiranno di identificare le caratteristiche da spostare.

Caratteristiche. Fare clic sulla caratteristica che si desidera spostare. Si noti che le caratteristiche saranno selezionate nell'anteprima della mappa. È anche possibile fare clic direttamente sulla caratteristica nell'anteprima della mappa.

X/Y. Queste caselle di testo elencano il punto centrale corrente della funzione nella mappa. Le unità sono visualizzate nelle coordinate della mappa. Può trattarsi di coordinate cartesiane locali (ad esempio lo State Plane Coordinate System americano) o di coordinate geografiche (dove la **X** è la longitudine e la **Y** la latitudine). Specificare le coordinate per la nuova posizione della caratteristica. È anche possibile fare clic su una caratteristica della mappa e trascinarla per spostarla. Le caselle di testo verranno aggiornate con la nuova posizione.

Elimina caratteristiche: È possibile eliminare dalla mappa le caratteristiche non desiderate. Può essere utile quando si desidera fare un po' di ordine eliminando alcune caratteristiche non di interesse nella visualizzazione della mappa.

Chiavi. Selezionare la chiave mappa contenente le etichette delle caratteristiche che consentiranno di identificare le caratteristiche da eliminare.

Caratteristiche. Fare clic sulla caratteristica che si desidera eliminare. Se si desidera eliminare più caratteristiche contemporaneamente, tenere premuto Ctrl mentre si fa clic sulle altre caratteristiche da eliminare. Si noti che le caratteristiche saranno selezionate anche nell'anteprima della mappa. È possibile fare clic e usare la combinazione Ctrl + clic sulle caratteristiche direttamente nell'anteprima della mappa, oltre a poterle selezionare dall'elenco.

Elimina singoli elementi: Oltre all'eliminazione di intere caratteristiche, è possibile eliminare i singoli elementi che compongono le caratteristiche, ad esempio laghi o isolotti. Questa opzione non è disponibile per le mappe di punti.

Elementi. Fare clic sugli elementi che si desidera eliminare. Se si desidera eliminare più elementi contemporaneamente, tenere premuto Ctrl mentre si fa clic sugli altri elementi da eliminare. Si noti che gli elementi saranno selezionati anche nell'anteprima della mappa. È possibile fare clic e usare la combinazione Ctrl + clic sugli elementi direttamente nell'anteprima della mappa, oltre a poterli selezionare dall'elenco. Poiché l'elenco dei nomi degli elementi non è descrittivo (ad ogni elemento è assegnato un numero all'interno della funzione), è necessario verificare la selezione nell'anteprima della mappa per verificare che siano stati selezionati gli elementi desiderati.

Imposta proiezione:

La proiezione della mappa specifica il modo in cui la Terra tridimensionale viene rappresentata in due dimensioni. Tutte le proiezioni causano distorsioni, tuttavia, alcune proiezioni sono più adatte a seconda che si stia visualizzando una mappa del globo o una mappa di una zona più specifica. Inoltre, alcune proiezioni mantengono la forma delle caratteristiche originarie. Tali proiezioni vengono chiamate proiezioni conformi. Questa opzione è disponibile solo per le mappe con coordinate geografiche (longitudine e latitudine).

A differenza di altre opzioni nel programma di utilità di conversione mappe, la proiezione può essere modificata dopo aver creato la visualizzazione della mappa.

Proiezione. Selezionare una proiezione di mappa. Se si sta creando una mappa del globo o di un emisfero, utilizzare le proiezioni *Locale*, *Mercatore* o *Winkel Tripel*. Per le aree più piccole, utilizzare le proiezioni *Locale*, *Conica conforme di Lambert* o *Trasversa di Mercatore*. Tutte le proiezioni utilizzano l'ellissoide WGS83 per il datum.

- La proiezione **Locale** viene sempre utilizzata quando la mappa è stata creata con un sistema di coordinate locali, ad esempio lo State Plane Coordinate System americano. Tali sistemi di coordinate sono definiti da coordinate cartesiane anziché da coordinate geografiche (longitudine e latitudine). Nella proiezione Locale, le linee orizzontali e verticali sono distanziate uniformemente in un sistema di coordinate cartesiane. La proiezione Locale non è conforme.
- La proiezione di **Mercatore** è una proiezione conforme per le mappe del globo. Le linee orizzontali e verticali sono linee rette sempre perpendicolari le une rispetto alle altre. Si noti che la proiezione di Mercatore si estende all'infinito man mano che ci si avvicina ai Poli, pertanto non può essere utilizzata se la mappa include il Polo Nord e il Polo Sud. Quando la mappa si avvicina a questi limiti, la distorsione è massima.
- La proiezione di **Winkel Tripel** è una proiezione non conforme per le mappe del globo. Benché non sia conforme, offre un buon compromesso tra forma e dimensione. Ad eccezione dell'Equatore e del Meridiano di Greenwich, tutte le linee sono curve. Se la propria mappa del globo include i Poli, questa proiezione è una buona scelta.
- Come si può dedurre dal nome, la proiezione **Conica conforme di Lambert** è una proiezione conforme e viene utilizzata per le mappe di masse continentali o masse terrestri di dimensioni minori che sono più estese verso est e ovest, piuttosto che verso nord e sud.
- La **Trasversa di Mercatore** è un'altra proiezione conforme utilizzata per le mappe di masse continentali o masse terrestri di dimensioni minori. Utilizzare questa proiezione per masse terrestri che sono più estese verso nord e sud, piuttosto che verso est e ovest.

Fase 4 - Fine

A questo punto è possibile aggiungere un commento per descrivere il file mappa e creare un file di dati di esempio dalle chiavi della mappa.

Chiavi mappa. Se sono presenti più chiavi nel file mappa, selezionare una chiave mappa di cui si desidera visualizzare le etichette delle caratteristiche nell'anteprima. Se si crea un file di dati dalla mappa, queste etichette verranno utilizzate per i valori dei dati.

Commento. Specificare un commento che descriva la mappa o che fornisca altre informazioni pertinenti ai propri utenti, ad esempio le sorgenti degli shapefile originali. Il commento apparirà nel sistema di gestione del Selezionatore modelli per lavagna grafica.

Crea un insieme di dati dalle etichette delle caratteristiche. Selezionare questa opzione se si desidera creare un file di dati dalle etichette delle funzioni visualizzate. Facendo clic su **Sfoggia...**, è possibile specificare un percorso ed un nome file. Se si aggiunge un'estensione *.txt*, il file verrà salvato come file di valori separati da tabulazioni. Se si aggiunge un'estensione *.csv*, il file verrà salvato come file di valori separati da virgola. Se si aggiunge l'estensione *.sav*, il file verrà salvato nel formato IBM SPSS Statistics. SAV è il valore predefinito quando non viene specificata alcuna estensione.

Distribuzione dei file mappa

Nella prima fase del programma di utilità di conversione mappe è stato scelto il percorso in cui salvare il file SMZ convertito. È possibile che in questa fase si sia anche scelto di aggiungere la mappa al sistema di gestione del Selezionatore modelli per lavagna grafica. In questo caso, la mappa sarà disponibile in qualsiasi prodotto IBM SPSS installato sullo stesso computer.

Per distribuire la mappa ad altri utenti, sarà necessario inviare loro il file SMZ. Questi utenti possono quindi utilizzare il sistema di gestione per importare la mappa. È sufficiente inviare il file il cui percorso è stato specificato nella fase 1. Se si desidera inviare un file che si trova nel sistema di gestione, è necessario prima esportarlo:

1. Nel Selezionatore modelli, fare clic su **Gestisci...**
2. Fare clic sulla scheda Mappa.
3. Selezionare la mappa da distribuire.
4. Fare clic su **Esporta...** e scegliere il percorso in cui si desidera salvare il file.

Ora è possibile inviare il file mappa fisico ad altri utenti. Gli utenti dovranno procedere al contrario e importare la mappa nel sistema di gestione.

Nodo Plot

I nodi Plot mostrano la relazione tra campi numerici. Per creare un grafico è possibile utilizzare punti (in modo da ottenere un grafico a dispersione) o linee. L'impostazione della Modalità X nella finestra di dialogo consente di creare tre tipi di grafici a linee.

Modalità X = Ordina

Se la Modalità X viene impostata su **Ordina**, i dati vengono ordinati in base ai valori per il campo rappresentati sull'asse x . Questa impostazione produce un'unica linea che attraversa il grafico da sinistra a destra. L'utilizzo di un campo nominale come sovrapposizione produce più linee di diverse tonalità che attraversano il grafico da sinistra a destra.

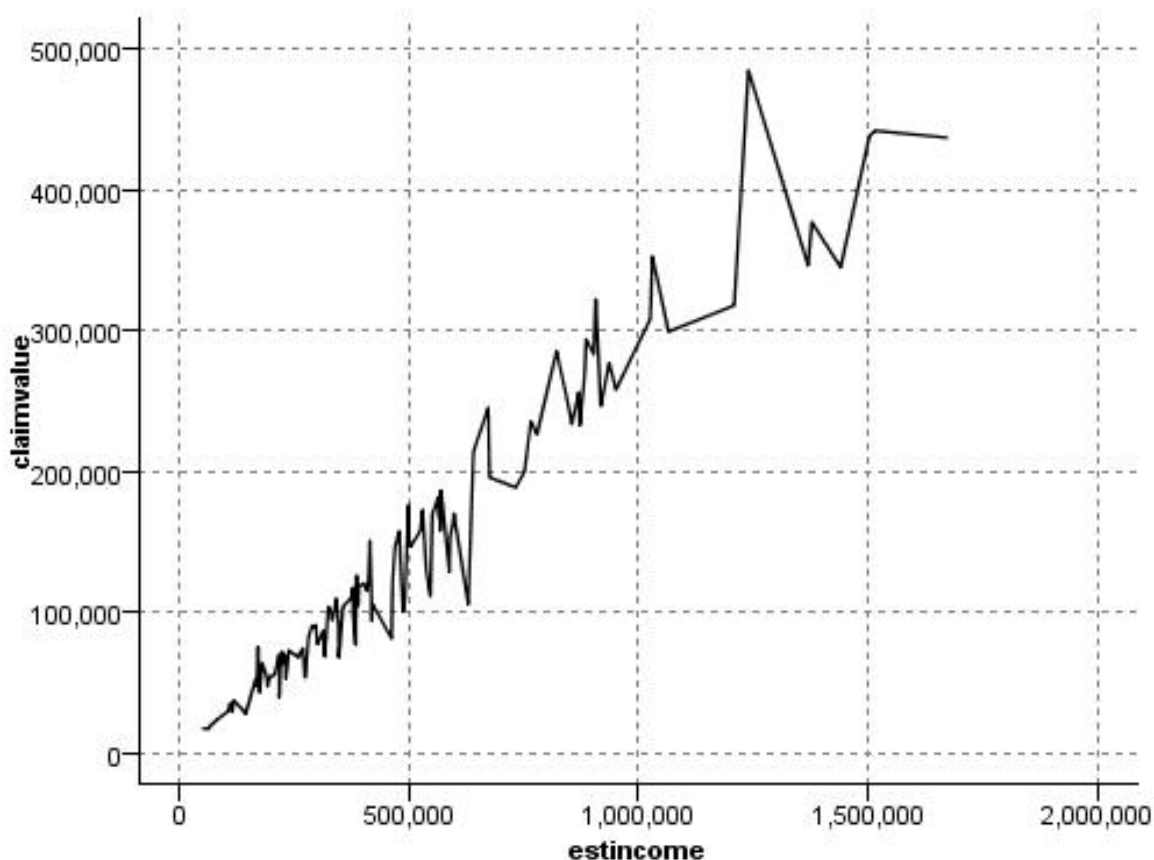


Figura 21. Grafico a linee con la Modalità X impostata su Ordina

Modalità X = Sovrapposizione

L'impostazione della Modalità X su **Sovrapposizione** crea più grafici a linee nello stesso grafico. I dati di un grafico di sovrapposizione non vengono ordinati. Se i valori sull'asse x aumentano, i dati verranno rappresentati su un'unica linea. Se i valori diminuiscono, viene aggiunta una nuova linea. Per esempio, quando x aumenta da 0 a 100 i valori y vengono rappresentati su un'unica linea. Quando x diminuisce fino a un valore inferiore a 100, nel grafico viene rappresentata una nuova linea in aggiunta alla linea esistente. La versione definitiva del grafico può includere diversi grafici utili per confrontare più serie di valori y . Questo tipo di grafico può essere utilizzato per dati con una componente temporale periodica, per esempio il consumo di energia elettrica in periodi successivi di 24 ore.

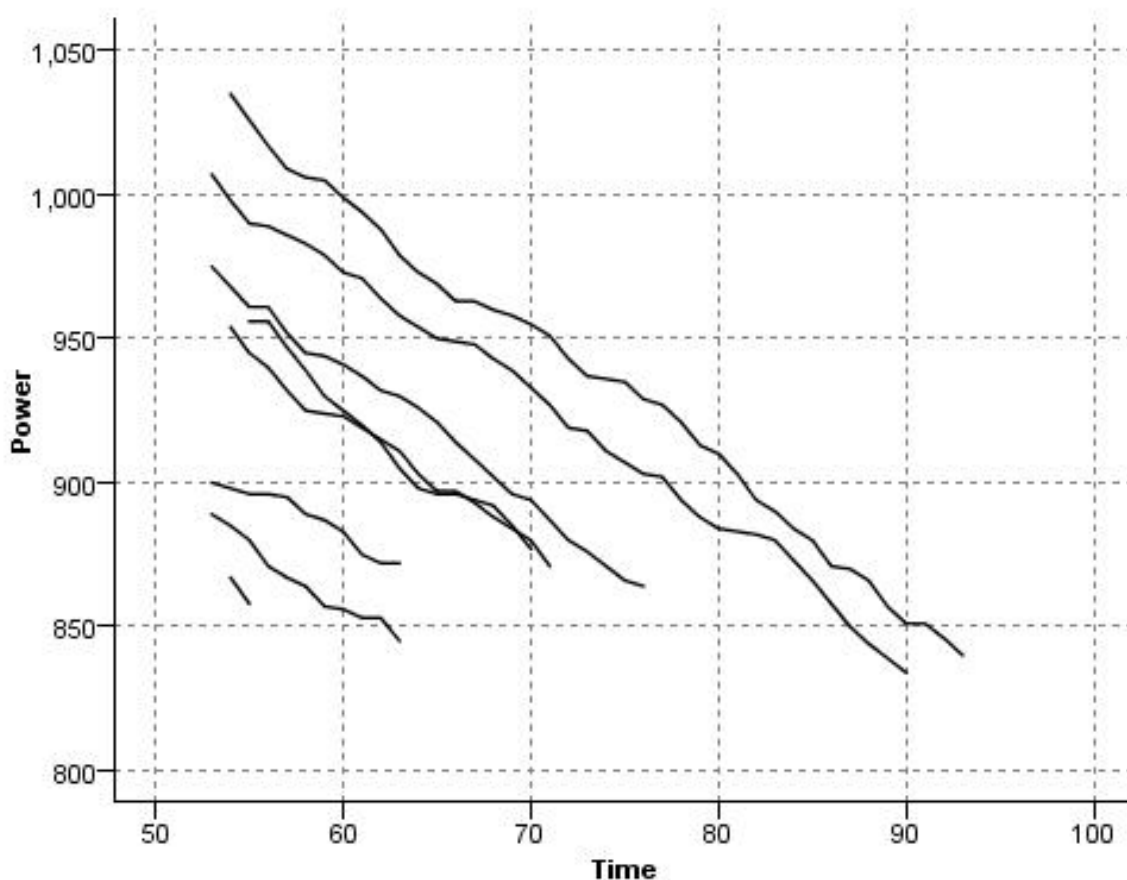


Figura 22. Grafico a linee con la Modalità X impostata su Sovrapposizione

Modalità X = Ordine di lettura

L'impostazione della Modalità X su **Ordine di lettura** consente di rappresentare i valori x e y così come vengono letti dalla sorgente dati. Questa opzione è utile per dati con una componente di serie temporali utilizzati per analizzare le tendenze o gli schemi che dipendono dall'ordine dei dati. Potrebbe essere necessario ordinare i dati prima di creare questo tipo di grafico. Inoltre, potrebbe essere utile confrontare due grafici simili con la Modalità X impostata su **Ordina** e su **Ordine di lettura**, in modo da stabilire fino a che punto uno schema dipenda dall'ordinamento.

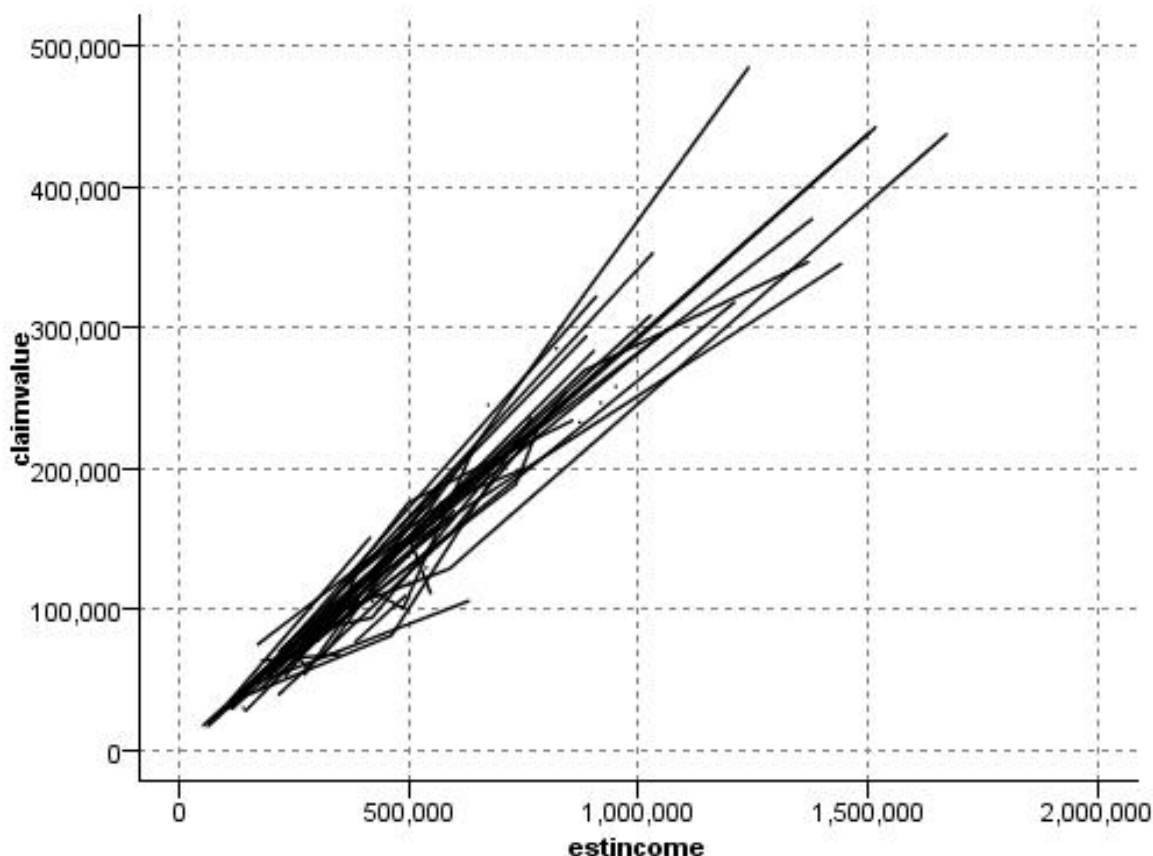


Figura 23. Grafico a linee mostrato in precedenza con l'impostazione Ordina, eseguito nuovamente con la Modalità X impostata su Ordine di lettura

Per generare grafici a dispersione e grafici a linee è anche possibile utilizzare il nodo Lavagna grafica. Tuttavia, il nodo qui illustrato dispone di un numero maggiore di opzioni da cui scegliere. Per ulteriori informazioni, consultare l'argomento "Tipi di visualizzazione della lavagna grafica incorporati disponibili" a pagina 176.

Scheda Nodo Plot

I nodi Plot mostrano il confronto tra i valori di un campo Y e i valori di un campo X. Spesso questi campi corrispondono rispettivamente a una variabile dipendente e a una variabile indipendente.

Campo X. Selezionare dall'elenco il campo da visualizzare lungo l'asse *x* orizzontale.

Campo Y. Selezionare dall'elenco il campo da visualizzare lungo l'asse *y* verticale.

Campo Z. Quando si fa clic sul pulsante del grafico 3-D è possibile scegliere dall'elenco un campo da visualizzare sull'asse *z*.

Sovrapposizione. Esistono diverse modalità di rappresentazione delle categorie per i valori dei dati. Ad esempio, è possibile utilizzare *raccolto_principale* come sovrapposizione di colore per indicare i valori *stima_reddito* e *valore_richiesta* per il raccolto principale coltivato da coloro che hanno presentato richiesta di indennizzo. Per ulteriori informazioni, consultare l'argomento "Effetti estetici, sovrapposizioni, riquadri e animazioni" a pagina 166.

Tipo sovrapposto. Specifica se viene visualizzata una funzione di sovrapposizione o di livellamento. Le funzioni di livellamento e sovrapposizione sono sempre calcolate come funzione di y.

- **Nessuna.** Non vengono visualizzate sovrapposizioni.
- **Livellamento.** Visualizza una curva di interpolazione calcolata con il metodo LOESS (locally-weighted iterative robust least squares regression, regressione iterativa di minimi quadrati ponderati localmente). Questo metodo calcola in modo efficiente una serie di regressioni, ciascuna concentrata su una piccola area del plot. In questo modo viene generata una serie di linee di regressione "locali" che vengono in seguito unite in modo da creare una curva.

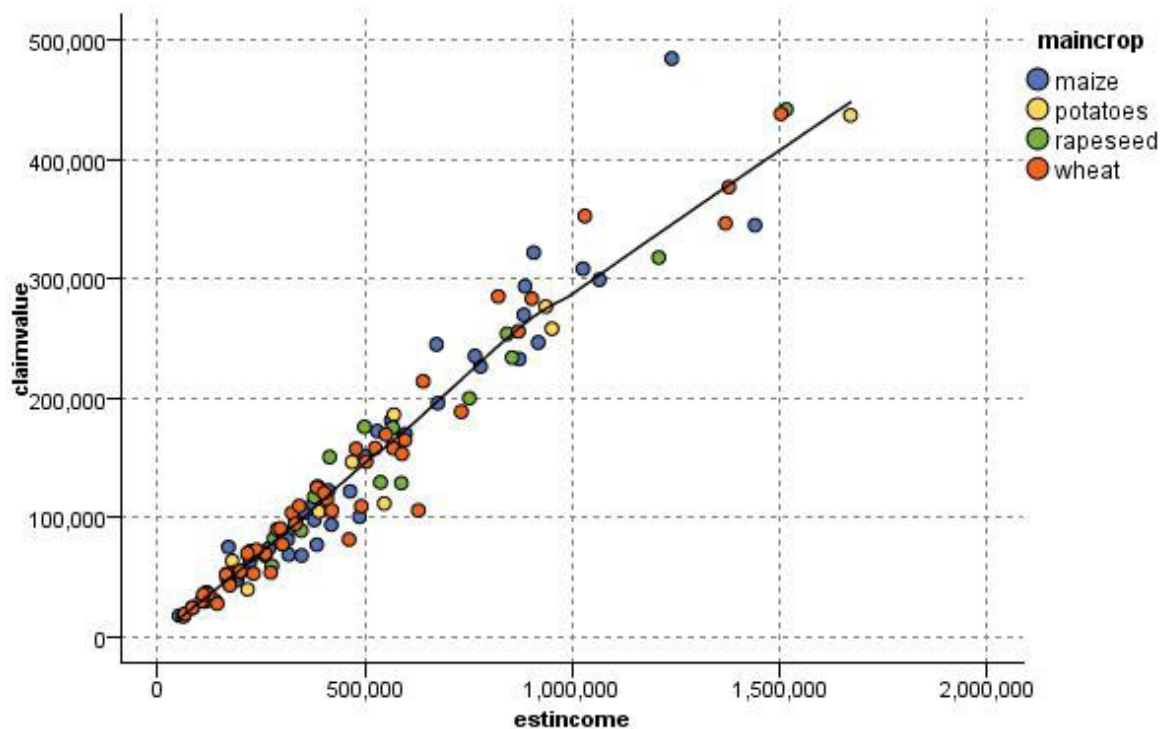


Figura 24. Plot con sovrapposizione del livellamento LOESS

- **Funzione.** Selezionare questa opzione per specificare una funzione conosciuta da confrontare con i valori effettivi. Ad esempio, per confrontare valori effettivi con valori previsti, è possibile rappresentare la funzione $y = x$ come una sovrapposizione. Specificare una funzione nella casella di testo $y =$. La funzione di default è $y = x$, tuttavia è possibile specificare una funzione qualsiasi, ad esempio una funzione quadratica o un'espressione arbitraria, al posto di x .

Nota: le funzioni sovrapposte non sono disponibili per un grafico contenente riquadri o animazione.

Dopo aver impostato le opzioni per un grafico, è possibile eseguire il grafico direttamente dalla finestra di dialogo facendo clic su **Esegui**. Tuttavia, è possibile che si desideri utilizzare la scheda Opzioni per impostare specifiche aggiuntive, ad esempio discretizzazione, Modalità X e stile.

Scheda Opzioni del nodo Plot

Stile. Selezionare **Punto** o **Linea** per lo stile del grafico. La selezione di **Linea** attiva il controllo **Modalità X**. Se si seleziona **Punto**, come forma di default del punto viene utilizzato il segno più (+). Dopo aver creato il grafico, è possibile modificare la forma del punto e la dimensione.

Modalità X. Per i grafici a linee è necessario selezionare un'opzione della Modalità X in modo da definire lo stile del grafico. Selezionare **Ordina**, **Sovrapposizione** oppure **Ordine di lettura**. Per **Sovrapposizione** o **Ordine di lettura**, è necessario specificare la dimensione dell'insieme di dati massima utilizzata per il campionamento dei primi n record. In caso contrario, verrà utilizzata l'impostazione di default di 2.000 record.

Intervallo X automatico. Selezionare questa opzione per utilizzare l'intero intervallo di valori nei dati lungo l'asse. Deselezionare questa opzione per utilizzare un sottoinsieme esplicito di valori compresi tra i valori **Min** e **Max** specificati. Digitare i valori direttamente oppure utilizzare le frecce. Gli intervalli automatici sono selezionati per impostazione predefinita per abilitare la creazione rapida dei grafici.

Intervallo Y automatico. Selezionare questa opzione per utilizzare l'intero intervallo di valori nei dati lungo l'asse. Deselezionare questa opzione per utilizzare un sottoinsieme esplicito di valori compresi tra i valori **Min** e **Max** specificati. Digitare i valori direttamente oppure utilizzare le frecce. Gli intervalli automatici sono selezionati per impostazione predefinita per abilitare la creazione rapida dei grafici.

Intervallo Z automatico. Solo quando nella scheda Plot è selezionato un grafico 3-D. Selezionare questa opzione per utilizzare l'intero intervallo di valori nei dati lungo l'asse. Deselezionare questa opzione per utilizzare un sottoinsieme esplicito di valori compresi tra i valori **Min** e **Max** specificati. Digitare i valori direttamente oppure utilizzare le frecce. Gli intervalli automatici sono selezionati per impostazione predefinita per abilitare la creazione rapida dei grafici.

Jitter. Anche noto come **agitazione**, è utile per grafici a punti di un insieme di dati in cui molti valori sono ripetuti. Per visualizzare più chiaramente la distribuzione dei valori, è possibile utilizzare l'opzione jitter per distribuire i punti in modo casuale attorno al valore effettivo.

Nota per gli utenti delle versioni precedenti di IBM SPSS Modeler: il valore di jitter usato in un grafico utilizza un tipo di misurazione diversa in questa versione di IBM SPSS Modeler. Nelle versioni precedenti il valore era un numero effettivo, ora è una percentuale della dimensione della cornice. Ciò significa che i valori di agitazione dei flussi delle versioni precedenti risulteranno probabilmente maggiori rispetto a quelli di questa versione. In questa versione, qualsiasi valore di agitazione diverso da zero verrà convertito nel valore 0,2.

Numero massimo di record per il plot. Specifica un metodo per rappresentare insiemi di dati di notevoli dimensioni. È possibile specificare una dimensione massima per l'insieme di dati oppure utilizzare l'impostazione di default di 2.000 record. Per gli insiemi di dati di grandi dimensioni le performance risultano migliori se si seleziona l'opzione **Bin** o **Campione**. In alternativa, è possibile rappresentare tutti i punti dati selezionando **Utilizza tutti i dati**, tuttavia occorre tenere presente che in questo modo le prestazioni del software potrebbero risultare notevolmente ridotte.

Nota: quando la Modalità X viene impostata su **Sovrapposizione** o su **Ordine di lettura**, queste opzioni non sono disponibili e vengono utilizzati solo i primi n record.

- **Bin.** Selezionare questa opzione per attivare gli intervalli quando l'insieme di dati contiene un numero di record maggiore di quello specificato. L'attivazione degli intervalli determina la divisione del grafico in griglie fini prima della rappresentazione effettiva e il conteggio del numero di punti che verrebbero visualizzati in ogni cella della griglia. Nel grafico finale viene rappresentato un unico punto per cella sul baricentro del bin (media di tutte le posizioni dei punti nel bin). La dimensione dei simboli rappresentati indica il numero di punti in quell'area (a meno che la dimensione non sia stata utilizzata come sovrapposizione). Grazie all'utilizzo del baricentro e della dimensione per la rappresentazione del numero di punti, il grafico a intervalli offre una modalità di rappresentazione eccellente per insiemi di dati di grandi dimensioni, in quanto impedisce la sovrarappresentazione in aree dense (masse di colore non differenziate) e riduce il rischio di rappresentazioni non valide (a causa di schemi di densità falsati). Le rappresentazioni non valide si verificano quando alcuni simboli (in particolare il segno più [+]) collidono in modo da produrre aree dense che non sono presenti nei dati grezzi.

- **Esempio.** Selezionare questa opzione per eseguire un campionamento casuale dei dati sul numero di record immesso nel campo testo. Il valore di default è 2,000.

Scheda Aspetto del nodo Plot

Prima di creare un grafico è possibile specificarne le opzioni dell'aspetto.

Titolo. Immettere il testo da utilizzare come titolo del grafico.

Sottotitolo. Immettere il testo da utilizzare come sottotitolo del grafico.

Didascalia. Immettere il testo che verrà utilizzato per la didascalia del grafico.

Etichetta X. Accettare l'etichetta dell'asse x (orizzontale) generata automaticamente oppure selezionare **Personalizzato** per specificare un'etichetta personalizzata.

Etichetta Y. Accettare l'etichetta dell'asse y (verticale) generata automaticamente oppure selezionare **Personalizzato** per specificare un'etichetta personalizzata.

Etichetta Z. Disponibile solo per i grafici 3-D. Accettare l'etichetta dell'asse z generata automaticamente oppure selezionare **Personalizzato** per specificare un'etichetta personalizzata.

Visualizza griglia. Questa opzione è selezionata per default e consente di visualizzare una griglia dietro il grafico per stabilire più facilmente i punti di interruzione di aree e sezioni. Le griglie sono sempre visualizzate in bianco. Sono di colore grigio solo se anche lo sfondo del grafico è bianco.

Utilizzo del grafico

I grafici ed i multiplot sono essenzialmente grafici di X rispetto a Y . Ad esempio, se si sta analizzando una potenziale frode nelle richieste di concessioni agricole (come illustrato in *fraud.str* nella cartella *Demos* dell'installazione IBM SPSS Modeler), è possibile che si desideri tracciare il reddito dichiarato nella richiesta rispetto a quello stimato da una rete neurale. Una sovrapposizione, ad esempio il tipo di raccolto, consentirà di vedere se esiste una relazione tra richieste (valore o numero) e tipo di raccolto.

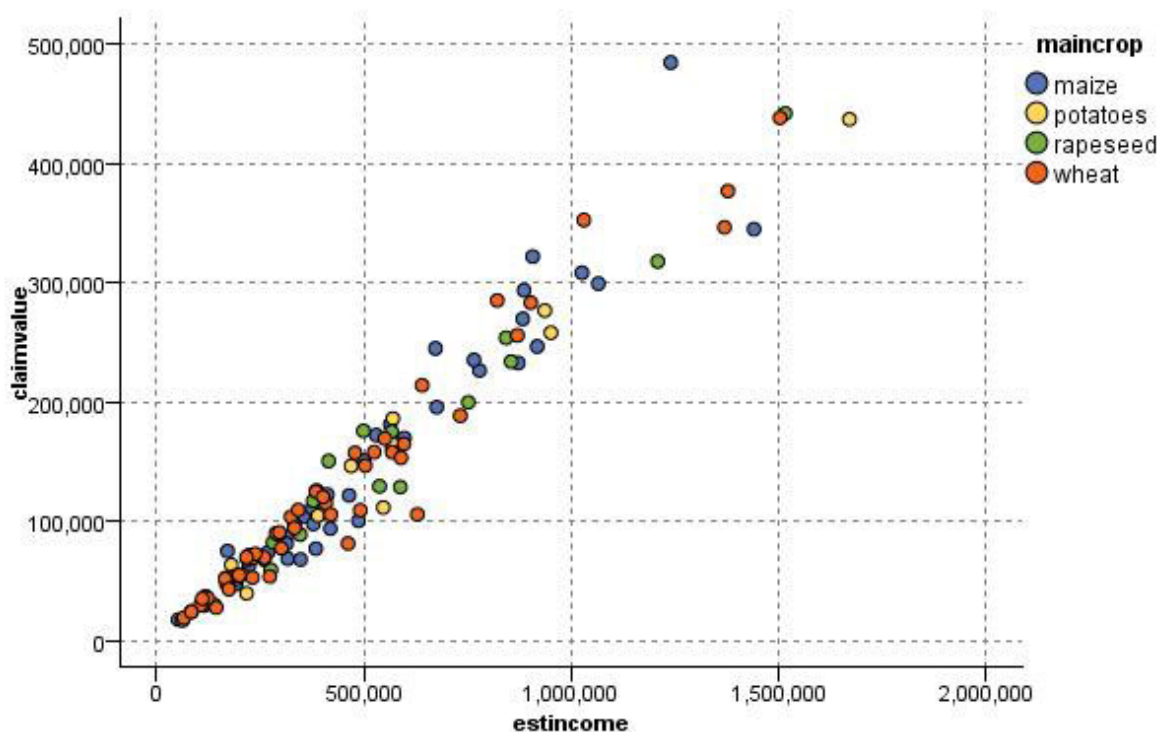


Figura 25. Grafico della relazione tra reddito stimato e valore della richiesta con il tipo raccolto principale come sovrapposizione

Dal momento che i grafici prodotti dai nodi Grafico, Multiplot e Valutazione sono visualizzazioni bidimensionali di Y rispetto a X , è facile interagire con essi selezionando le aree, contrassegnando elementi o persino disegnando sezioni mediante il mouse. È inoltre possibile generare nodi per i dati rappresentati da queste aree, sezioni o elementi. Per ulteriori informazioni, consultare l'argomento "Esplorazione dei grafici" a pagina 243.

Nodo distribuzione

Un grafico o tabella di distribuzione mostra l'occorrenza di valori simbolici (non numerici), per esempio tipo di ipoteca o sesso, in un insieme di dati. Un nodo distribuzione viene in genere utilizzato per mostrare gli squilibri nei dati che possono essere corretti se si aggiunge un nodo bilanciamento prima di creare un modello. Per generare automaticamente un nodo bilanciamento, utilizzare il menu Genera nella finestra di un grafico o tabella di distribuzione.

Per generare grafici con barre dei conteggi è anche possibile utilizzare il nodo Lavagna grafica. Tuttavia, il nodo qui illustrato dispone di un numero maggiore di opzioni da cui scegliere. Per ulteriori informazioni, consultare l'argomento "Tipi di visualizzazione della lavagna grafica incorporati disponibili" a pagina 176.

Nota: per visualizzare l'occorrenza di valori numerici, è necessario utilizzare un nodo Istogramma.

Scheda Plot del nodo distribuzione

Grafico. Selezionare il tipo di distribuzione. Selezionare **Campi selezionati** per visualizzare la distribuzione del campo selezionato. Selezionare **Tutti i flag (valori veri)** per visualizzare la distribuzione dei valori veri per i campi flag nell'insieme di dati.

Campo. Selezionare un campo nominale o flag per cui mostrare la distribuzione dei valori. Nell'elenco sono inclusi solo i campi non esplicitamente impostati come numerici.

Sovrapposizione. Selezionare un campo nominale o flag da utilizzare come sovrapposizione di colore, che illustra la distribuzione dei propri valori all'interno di ciascun valore del campo specificato. Ad esempio, è possibile utilizzare la risposta a una campagna di marketing (*pep*) come sovrapposizione per numero di figli (*figli*) per rappresentare la risposta in base alla dimensione del nucleo familiare. Per ulteriori informazioni, consultare l'argomento "Effetti estetici, sovrapposizioni, riquadri e animazioni" a pagina 166.

Normalizza in base al colore. Selezionare questa opzione per ridimensionare le barre in modo che tutte le barre occupino l'intera larghezza del grafico. I valori di sovrapposizione corrispondono a una proporzione di ogni barra, pertanto il confronto tra le categorie risulta più semplice.

Ordina. Selezionare il metodo utilizzato per visualizzare i valori nel grafico di distribuzione. Selezionare **Alfabetico** per utilizzare l'ordine alfabetico, oppure **Per conteggio** per elencare i valori in ordine di occorrenza decrescente.

Scala proporzionale. Selezionare per ridimensionare la distribuzione dei valori in modo che il valore con il conteggio maggiore occupi l'intera larghezza del grafico. Tutte le altre barre vengono graduate rispetto a questo valore. Se questa opzione è deselezionata, le barre vengono graduate in base al conteggio totale per ogni valore.

Scheda Aspetto del nodo distribuzione

Prima di creare un grafico è possibile specificarne le opzioni dell'aspetto.

Titolo. Immettere il testo da utilizzare come titolo del grafico.

Sottotitolo. Immettere il testo da utilizzare come sottotitolo del grafico.

Didascalia. Immettere il testo che verrà utilizzato per la didascalia del grafico.

Etichetta X. Accettare l'etichetta dell'asse x (orizzontale) generata automaticamente oppure selezionare **Personalizzato** per specificare un'etichetta personalizzata.

Etichetta Y. Accettare l'etichetta dell'asse y (verticale) generata automaticamente oppure selezionare **Personalizzato** per specificare un'etichetta personalizzata.

Visualizza griglia. Questa opzione è selezionata per default e consente di visualizzare una griglia dietro il grafico per stabilire più facilmente i punti di interruzione di aree e sezioni. Le griglie sono sempre visualizzate in bianco. Sono di colore grigio solo se anche lo sfondo del grafico è bianco.

Utilizzo di un nodo distribuzione

I nodi distribuzione vengono utilizzati per rappresentare la distribuzione dei valori simbolici in un insieme di dati. Vengono spesso aggiunti prima dei nodi di manipolazione, in modo da analizzare i dati e correggere gli eventuali squilibri. Ad esempio, se nell'ambito di dati relativi a clienti che hanno risposto favorevolmente a campagne di marketing le istanze di famiglie con figli sono più frequenti rispetto ad altre tipologie di clienti, è possibile che si desideri ridurre tali istanze in modo da generare una regola più utile nelle successive operazioni di data mining. Un nodo distribuzione consente di analizzare questo tipo di squilibri e prendere decisioni appropriate.

Il nodo distribuzione è insolito in quanto produce sia un grafico sia una tabella per analizzare i dati.

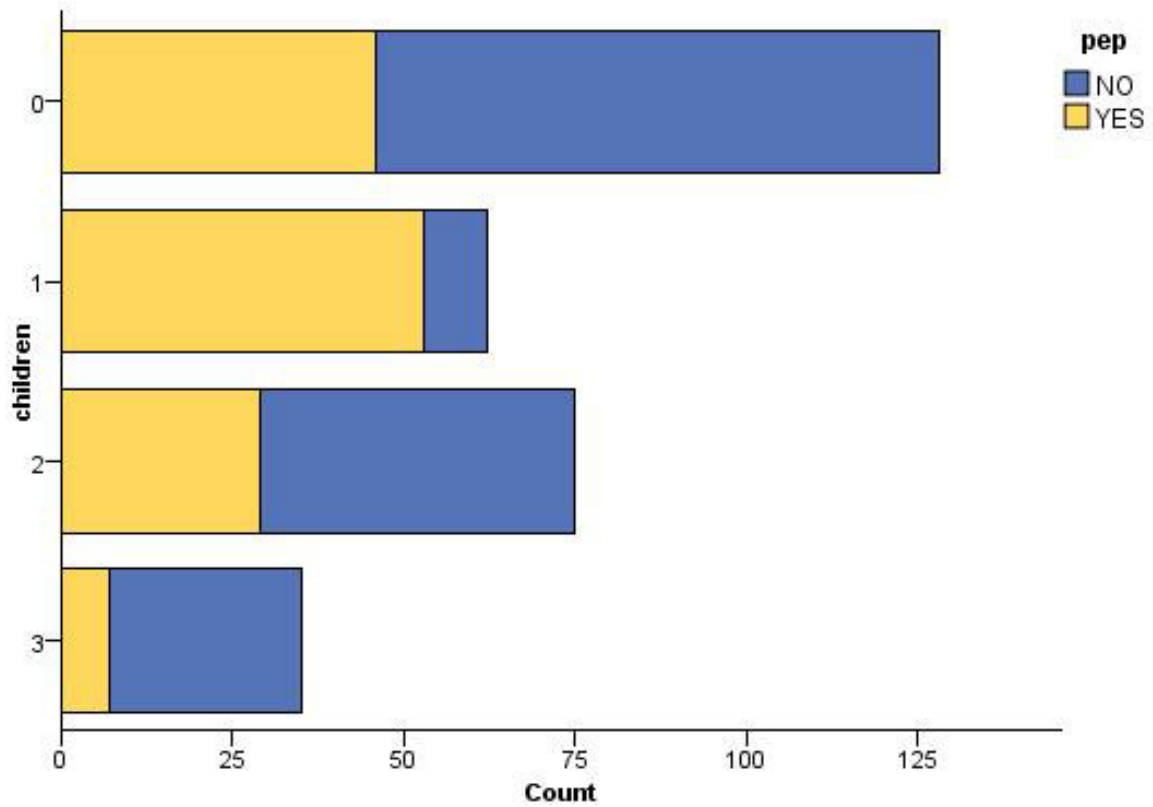


Figura 26. Grafico di distribuzione che mostra il numero di individui con o senza figli che hanno risposto favorevolmente a una campagna di marketing

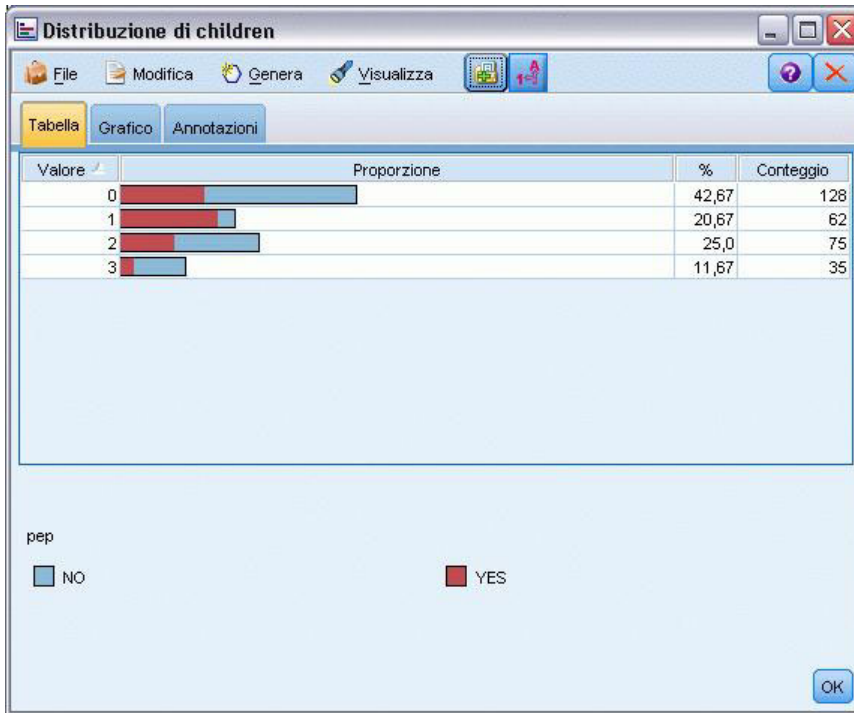


Figura 27. Tabella di distribuzione che mostra la proporzione degli individui con o senza figli che hanno risposto favorevolmente a una campagna di marketing

Dopo aver creato un grafico e una tabella di distribuzione e averne analizzato i risultati, è possibile utilizzare le opzioni dei menu per raggruppare i valori, copiarli e generare una serie di nodi per preparare i dati. Inoltre, è possibile copiare o esportare le informazioni della tabella o del grafico per utilizzarle in altre applicazioni, come MS Word o MS PowerPoint. Per ulteriori informazioni, consultare l'argomento "Stampa, salvataggio, copia ed esportazione di grafici" a pagina 268.

Per selezionare e copiare valori da una tabella di distribuzione

1. Fare clic e tenere premuto il pulsante del mouse mentre si trascina il cursore sulle righe per selezionare un insieme di valori. È anche possibile utilizzare il menu Modifica e scegliere **Seleziona tutto** per selezionare tutti i valori.
2. Dal menu Modifica scegliere **Copia tabella** o **Copia tabella (inc. nomi di campo)**.
3. Incollare negli Appunti o nell'applicazione desiderata.

Nota: le barre non vengono copiate direttamente. I valori della tabella, invece, vengono copiati. Pertanto, i valori sovrapposti non verranno visualizzati nella tabella copiata.

Per raggruppare valori da una tabella di distribuzione

1. Selezionare i valori da raggruppare utilizzando il metodo Ctrl+clic.
2. Dal menu Modifica scegliere **Raggruppa**.

Nota: quando si raggruppano e si dividono i valori, il grafico nella scheda Grafico viene automaticamente ridisegnato per visualizzare le modifiche.

È anche possibile:

- Dividere valori selezionando il nome del gruppo nell'elenco di distribuzione e scegliendo **Separa** dal menu Modifica.
- Modificare gruppi selezionando il nome del gruppo nell'elenco di distribuzione e scegliendo **Modifica gruppo** dal menu Modifica. In questo modo viene aperta una finestra di dialogo in cui è possibile modificare i valori del gruppo.

Opzioni del menu Genera

È possibile utilizzare le opzioni del menu Genera per selezionare un sottoinsieme di dati, derivare un campo flag, raggruppare nuovamente valori, riclassificare valori o bilanciare i dati di un grafico o di una tabella. Tali opzioni generano un nodo Data Preparation e lo posizionano nell'area del flusso. Per utilizzare il nodo generato è necessario connetterlo a un flusso esistente. Per ulteriori informazioni, consultare l'argomento "Generazioni di nodi da grafici" a pagina 251.

Nodo Istogramma

I nodi Istogramma mostrano l'occorrenza dei valori per i campi numerici. Vengono spesso utilizzati per analizzare i dati prima delle manipolazioni e della creazione del modello. Come il nodo distribuzione, anche i nodi Istogramma vengono frequentemente utilizzati per rivelare squilibri nei dati. Benché per generare un istogramma sia possibile utilizzare anche il nodo Lavagna grafica, il nodo qui illustrato dispone di un numero maggiore di opzioni tra cui scegliere. Per ulteriori informazioni, consultare l'argomento "Tipi di visualizzazione della lavagna grafica incorporati disponibili" a pagina 176.

Nota: per mostrare l'occorrenza di valori per campi simbolici, è necessario utilizzare un nodo distribuzione.

Scheda Plot del nodo Istogramma

Campo. Selezionare un campo numerico per cui mostrare la distribuzione dei valori. Verranno elencati solo i campi non esplicitamente definiti come simbolici (categoriali).

Sovrapposizione. Selezionare un campo simbolico per visualizzare le categorie dei valori per il campo specificato. La selezione di un campo di sovrapposizione converte l'istogramma in un grafico sovrapposto in cui le varie categorie del campo di sovrapposizione sono rappresentate da colori diversi. Con il nodo Istogrammi sono disponibili tre tipi di sovrapposizioni: colore, riquadro e animazione. Per ulteriori informazioni, consultare l'argomento "Effetti estetici, sovrapposizioni, riquadri e animazioni" a pagina 166.

Scheda Opzioni del nodo Istogramma

Intervallo X automatico. Selezionare questa opzione per utilizzare l'intero intervallo di valori nei dati lungo l'asse. Deselezionare questa opzione per utilizzare un sottoinsieme esplicito di valori compresi tra i valori **Min** e **Max** specificati. Digitare i valori direttamente oppure utilizzare le frecce. Gli intervalli automatici sono selezionati per impostazione predefinita per abilitare la creazione rapida dei grafici.

Bin. Selezionare **Per numero** oppure **Per larghezza**.

- Selezionare **Per numero** per visualizzare un numero fisso di barre la cui larghezza dipende dall'intervallo e dal numero di bin specificati. Indicare il numero di bin da utilizzare nel grafico nell'opzione **N. di bin**. Utilizzare le frecce per modificare il numero.
- Selezionare **Per larghezza** per creare un grafico con barre di larghezza fissa. Il numero di bin dipende dalla larghezza e dall'intervallo di valori specificati. Indicare la larghezza delle barre nell'opzione **Larghezza bin**.

Normalizza in base al colore. Selezionare questa opzione per impostare tutte le barre sulla stessa altezza, visualizzando i valori sovrapposti come percentuale dei casi totali in ogni barra.

Mostra curva normale. Selezionare questa opzione per aggiungere al grafico una curva normale che mostri la media e la varianza dei dati.

Sezioni separate per ogni colore. Selezionare questa opzione per visualizzare ogni valore sovrapposto come una sezione separata nel grafico.

Scheda Aspetto del nodo Istogramma

Prima di creare un grafico è possibile specificarne le opzioni dell'aspetto.

Titolo. Immettere il testo da utilizzare come titolo del grafico.

Sottotitolo. Immettere il testo da utilizzare come sottotitolo del grafico.

Didascalia. Immettere il testo che verrà utilizzato per la didascalia del grafico.

Etichetta X. Accettare l'etichetta dell'asse x (orizzontale) generata automaticamente oppure selezionare **Personalizzato** per specificare un'etichetta personalizzata.

Etichetta Y. Accettare l'etichetta dell'asse y (verticale) generata automaticamente oppure selezionare **Personalizzato** per specificare un'etichetta personalizzata.

Visualizza griglia. Questa opzione è selezionata per default e consente di visualizzare una griglia dietro il grafico per stabilire più facilmente i punti di interruzione di aree e sezioni. Le griglie sono sempre visualizzate in bianco. Sono di colore grigio solo se anche lo sfondo del grafico è bianco.

Utilizzo degli istogrammi

Gli istogrammi mostrano la distribuzione dei valori in un campo numerico i cui valori sono disposti lungo l'asse x . Gli istogrammi funzionano in modo simile ai grafici raccolta. Le raccolte mostrano la distribuzione dei valori di un campo numerico *rispetto ai valori di un altro campo* anziché l'occorrenza dei valori per un singolo campo.

Dopo avere creato un grafico è possibile esaminare i risultati e definire delle sezioni per suddividere i valori sull'asse x o definire delle aree. È possibile anche contrassegnare degli elementi del grafico. Per ulteriori informazioni, consultare l'argomento "Esplorazione dei grafici" a pagina 243.

È possibile utilizzare le opzioni del menu Genera per creare nodi bilanciamento, seleziona o nuovo campo con i dati del grafico o più specificamente all'interno di sezioni, aree o elementi contrassegnati. Questo tipo di grafico viene spesso utilizzato prima dei nodi di manipolazione per analizzare i dati e correggere eventuali squilibri mediante la generazione di un nodo bilanciamento da utilizzare nel flusso a partire dal grafico. È anche possibile generare un nodo Ricava indicatore per aggiungere un campo che indichi la sezione in cui ogni record è incluso, oppure un nodo Seleziona per selezionare tutti i record entro un particolare insieme o intervallo di valori. Queste operazioni consentono di delimitare in modo preciso un determinato sottoinsieme di dati per eseguire ulteriori esplorazioni. Per ulteriori informazioni, consultare l'argomento "Generazioni di nodi da grafici" a pagina 251.

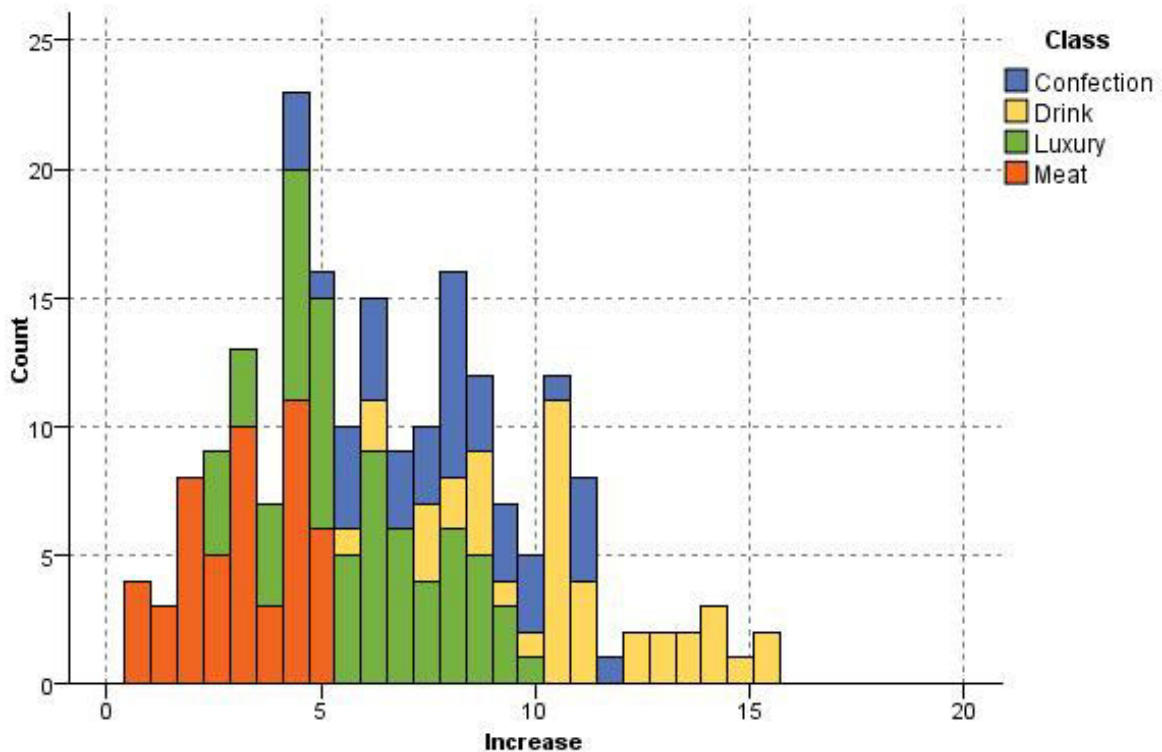


Figura 28. Istogramma che mostra la distribuzione della crescita degli acquisti per categoria a seguito di una promozione

Nodo Raccolta

Le raccolte sono simili agli istogrammi tranne per il fatto che le prime mostrano la distribuzione dei valori di un campo numerico rispetto ai valori di un altro campo anziché l'occorrenza dei valori per un singolo campo. Una raccolta è utile per illustrare una variabile o un campo i cui valori vengono modificati nel tempo. La grafica 3-D consente inoltre di includere un asse simbolico che visualizza le distribuzioni per categoria. Le raccolte bidimensionali sono visualizzate come grafici a barre sovrapposte con sovrapposizioni (se utilizzate). Per ulteriori informazioni, consultare l'argomento "Effetti estetici, sovrapposizioni, riquadri e animazioni" a pagina 166.

Scheda Plot del nodo Raccolta

Riassumi. Selezionare un campo i cui valori verranno raccolti e visualizzati in base all'intervallo di valori del campo specificato in **Su**. Vengono elencati solo i campi che non sono stati definiti come simbolici.

Su. Selezionare un campo i cui valori verranno utilizzati per visualizzare il campo specificato in **Riassumi**.

Per. Abilitata durante la creazione di un grafico 3D, questa opzione consente di selezionare un campo nominale o flag utilizzato per visualizzare il campo di raccolta per categorie.

Operazione. Selezionare il tipo di operazione che verrà rappresentata da ogni barra nel grafico raccolta. Le opzioni includono **Somma**, **Media**, **Max**, **Min** e **Deviazione standard**.

Sovrapposizione. Selezionare un campo simbolico per visualizzare le categorie dei valori per il campo selezionato. Se viene selezionato un campo di sovrapposizione, il grafico raccolta viene convertito e vengono aggiunte più barre di colori diversi per ogni categoria. Con questo nodo sono disponibili tre tipi di sovrapposizioni: colore, riquadro e animazione. Per ulteriori informazioni, consultare l'argomento "Effetti estetici, sovrapposizioni, riquadri e animazioni" a pagina 166.

Scheda Opzioni del nodo Raccolta

Intervallo X automatico. Selezionare questa opzione per utilizzare l'intero intervallo di valori nei dati lungo l'asse. Deselezionare questa opzione per utilizzare un sottoinsieme esplicito di valori compresi tra i valori **Min** e **Max** specificati. Digitare i valori direttamente oppure utilizzare le frecce. Gli intervalli automatici sono selezionati per impostazione predefinita per abilitare la creazione rapida dei grafici.

Bin. Selezionare **Per numero** oppure **Per larghezza**.

- Selezionare **Per numero** per visualizzare un numero fisso di barre la cui larghezza dipende dall'intervallo e dal numero di bin specificati. Indicare il numero di bin da utilizzare nel grafico nell'opzione **N. di bin**. Utilizzare le frecce per modificare il numero.
- Selezionare **Per larghezza** per creare un grafico con barre di larghezza fissa. Il numero di bin dipende dalla larghezza e dall'intervallo di valori specificati. Indicare la larghezza delle barre nell'opzione **Larghezza bin**.

Scheda Aspetto del nodo Raccolta

Prima di creare un grafico è possibile specificarne le opzioni dell'aspetto.

Titolo. Immettere il testo da utilizzare come titolo del grafico.

Sottotitolo. Immettere il testo da utilizzare come sottotitolo del grafico.

Didascalia. Immettere il testo che verrà utilizzato per la didascalia del grafico.

Su. Accettare l'etichetta automaticamente generata oppure selezionare **Personalizzato** per specificare un'etichetta.

Riassumi. Accettare l'etichetta automaticamente generata oppure selezionare **Personalizzato** per specificare un'etichetta.

Per. Accettare l'etichetta automaticamente generata oppure selezionare **Personalizzato** per specificare un'etichetta.

Visualizza griglia. Questa opzione è selezionata per default e consente di visualizzare una griglia dietro il grafico per stabilire più facilmente i punti di interruzione di aree e sezioni. Le griglie sono sempre visualizzate in bianco. Sono di colore grigio solo se anche lo sfondo del grafico è bianco.

Il seguente esempio mostra dove vengono posizionate le opzioni dell'aspetto su una versione 3-D del grafico.

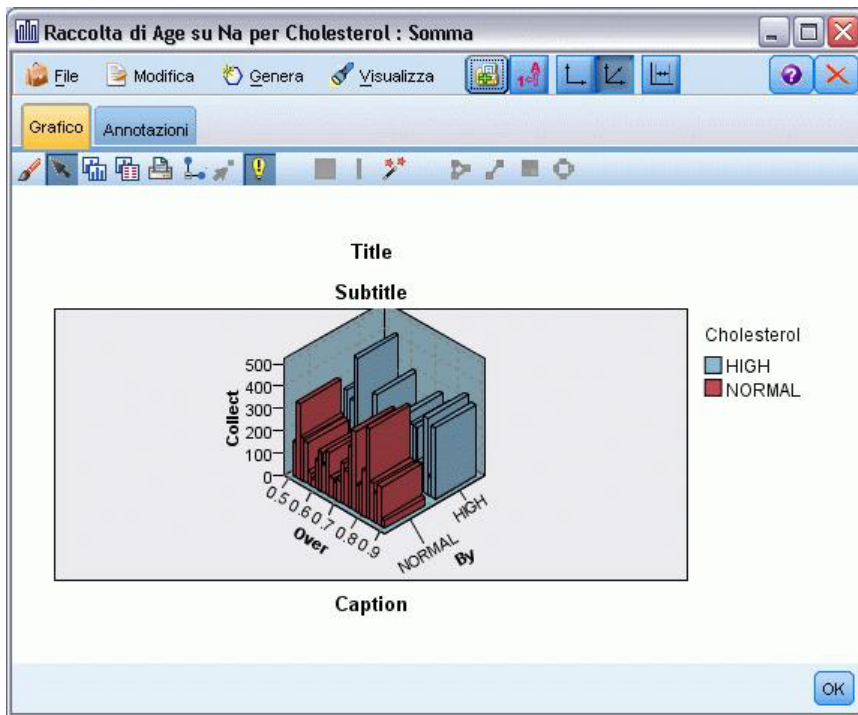


Figura 29. Posizione delle opzioni dell'aspetto in un grafico raccolta 3-D

Utilizzo del grafico raccolta

Le raccolte mostrano la distribuzione dei valori di un campo numerico *rispetto ai valori di un altro campo* anziché l'occorrenza dei valori per un singolo campo. Gli istogrammi funzionano in modo simile ai grafici raccolta. Gli istogrammi mostrano la distribuzione dei valori in un campo numerico i cui valori sono disposti lungo l'asse x .

Dopo avere creato un grafico è possibile esaminare i risultati e definire delle sezioni per suddividere i valori sull'asse x o definire delle aree. È possibile anche contrassegnare degli elementi del grafico. Per ulteriori informazioni, consultare l'argomento "Esplorazione dei grafici" a pagina 243.

È possibile utilizzare le opzioni del menu Genera per creare nodi bilanciamento, seleziona o nuovo campo con i dati del grafico o più specificamente all'interno di sezioni, aree o elementi contrassegnati. Questo tipo di grafico viene spesso utilizzato prima dei nodi di manipolazione per analizzare i dati e correggere eventuali squilibri mediante la generazione di un nodo bilanciamento da utilizzare nel flusso a partire dal grafico. È anche possibile generare un nodo Ricava indicatore per aggiungere un campo che indichi la sezione in cui ogni record è incluso, oppure un nodo Seleziona per selezionare tutti i record entro un particolare insieme o intervallo di valori. Queste operazioni consentono di delimitare in modo preciso un determinato sottoinsieme di dati per eseguire ulteriori esplorazioni. Per ulteriori informazioni, consultare l'argomento "Generazioni di nodi da grafici" a pagina 251.

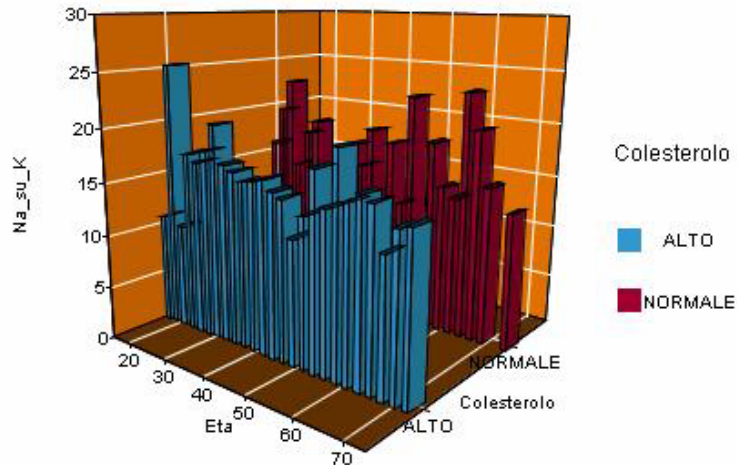


Figura 30. Grafico raccolta 3-D che mostra la somma di Na_su_K rispetto a Età per livelli di colesterolo alti e nella norma

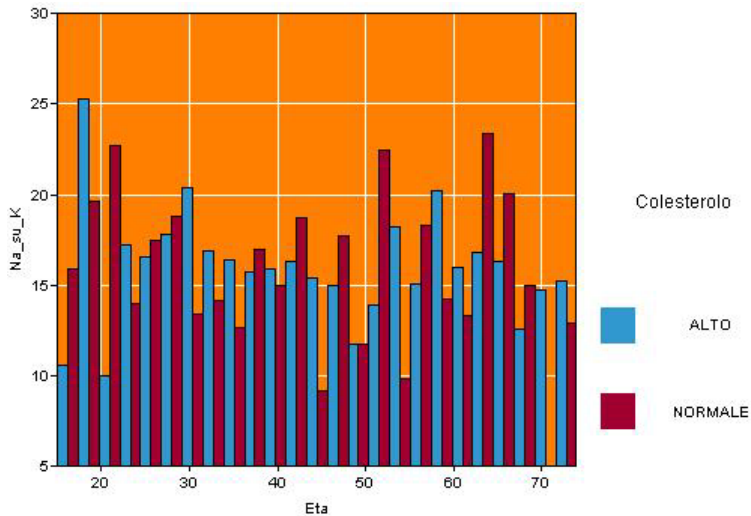


Figura 31. Grafico raccolta senza la visualizzazione dell'asse z ma con il campo Colesterolo come sovrapposizione di colore

Nodo Multiplot

Un multiplot è un tipo speciale di grafico che consente di visualizzare più campi Y in un singolo campo X. I campi Y sono rappresentati come linee colorate e ognuno di essi equivale a un nodo Plot con lo Stile impostato su **Linea** e la Modalità X impostata su **Ordina**. I multiplot sono utili quando si utilizzano dati di serie temporali e si desidera analizzare la fluttuazione di più variabili nel tempo.

Scheda Grafico di un nodo Multiplot

Campo X. Selezionare dall'elenco il campo da visualizzare lungo l'asse x orizzontale.

Campi Y. Selezionare dall'elenco uno o più campi da visualizzare sull'intervallo dei valori del campo X. Utilizzare il pulsante Selettore di campo per selezionare più campi. Fare clic sul pulsante contrassegnato da una X per rimuovere i campi dall'elenco.

Sovrapposizione. Esistono diverse modalità di rappresentazione delle categorie per i valori dei dati. Ad esempio, è possibile utilizzare una sovrapposizione di animazione per visualizzare più grafici per ogni valore nei dati. Questo è utile per gli insiemi che contengono più di 10 categorie. Nel caso di insiemi con più di 15 categorie si può notare una diminuzione delle prestazioni. Per ulteriori informazioni, consultare l'argomento "Effetti estetici, sovrapposizioni, riquadri e animazioni" a pagina 166.

Normalizza. Selezionare per ridimensionare tutti i valori Y in base all'intervallo 0–1 per la visualizzazione nel grafico. La normalizzazione consente di visualizzare in modo chiaro la relazione tra linee che potrebbero risultare non completamente visibili a causa di differenze nell'intervallo dei valori di ogni serie ed è consigliata quando si rappresentano più linee nello stesso grafico o quando si confrontano plot in riquadri affiancati (la normalizzazione non è necessaria quando tutti i valori di dati rientrano in un intervallo simile).

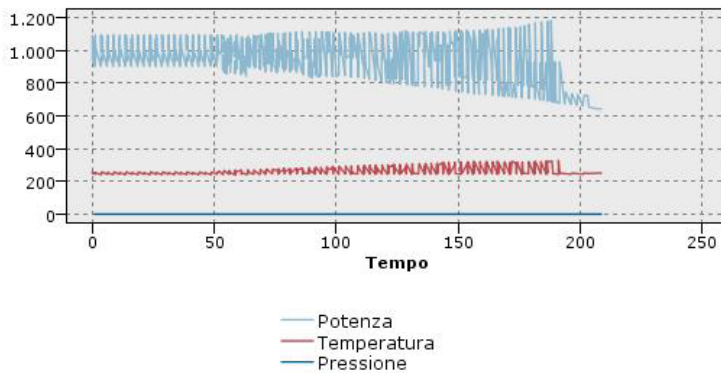


Figura 32. Multiplot standard che mostra la fluttuazione di una centrale elettrica nel tempo (si noti che senza la normalizzazione è impossibile vedere il grafico relativo alla pressione)

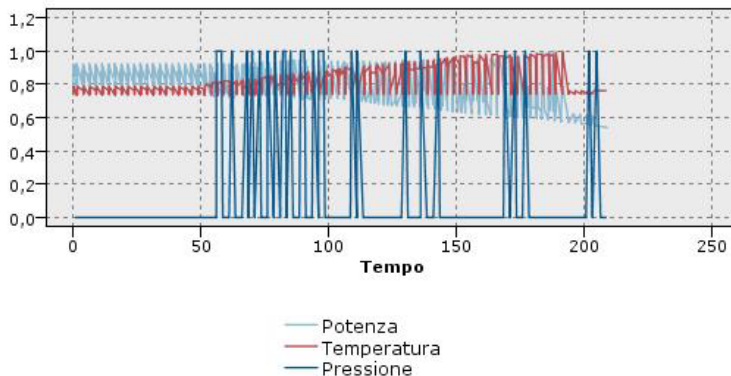


Figura 33. Multiplot normalizzato che mostra un grafico relativo a valori di pressione

Funzione sovrapposta. Selezionare questa opzione per specificare una funzione conosciuta da confrontare con i valori effettivi. Ad esempio, per confrontare valori effettivi con valori previsti, è possibile rappresentare la funzione $y = x$ come una sovrapposizione. Specificare una funzione nella casella di testo $y =$. La funzione di default è $y = x$, tuttavia è possibile specificare una funzione qualsiasi, ad esempio una funzione quadratica o un'espressione arbitraria, al posto di x .

Nota: le funzioni sovrapposte non sono disponibili per un grafico contenente riquadri o animazione.

Quando il numero di record è maggiore di. Specifica un metodo per rappresentare insiemi di dati di notevoli dimensioni. È possibile specificare una dimensione massima per l'insieme di dati oppure

utilizzare l'impostazione di default di 2.000 punti. Per gli insiemi di dati di grandi dimensioni le performance risultano migliori se si seleziona l'opzione **Bin** o **Campione**. In alternativa, è possibile rappresentare tutti i punti dati selezionando **Utilizza tutti i dati**, tuttavia occorre tenere presente che in questo modo le prestazioni del software potrebbero risultare notevolmente ridotte.

Nota: quando la Modalità X viene impostata su **Sovrapposizione** o su **Ordine di lettura**, queste opzioni non sono disponibili e vengono utilizzati solo i primi n record.

- **Bin.** Selezionare questa opzione per attivare gli intervalli quando l'insieme di dati contiene un numero di record maggiore di quello specificato. L'attivazione degli intervalli determina la divisione del grafico in griglie fini prima della rappresentazione effettiva e il conteggio del numero di connessioni che verrebbero visualizzate in ogni cella della griglia. Nel grafico finale viene rappresentata un'unica connessione per cella sul baricentro del bin (media di tutti i punti di connessione nel bin).
- **Esempio.** Selezionare questa opzione per eseguire un campionamento casuale dei dati sul numero di record specificato.

Scheda Aspetto del nodo Multiplot

Prima di creare un grafico è possibile specificarne le opzioni dell'aspetto.

Titolo. Immettere il testo da utilizzare come titolo del grafico.

Sottotitolo. Immettere il testo da utilizzare come sottotitolo del grafico.

Didascalia. Immettere il testo che verrà utilizzato per la didascalia del grafico.

Etichetta X. Accettare l'etichetta dell'asse x (orizzontale) generata automaticamente oppure selezionare **Personalizzato** per specificare un'etichetta personalizzata.

Etichetta Y. Accettare l'etichetta dell'asse y (verticale) generata automaticamente oppure selezionare **Personalizzato** per specificare un'etichetta personalizzata.

Visualizza griglia. Questa opzione è selezionata per default e consente di visualizzare una griglia dietro il grafico per stabilire più facilmente i punti di interruzione di aree e sezioni. Le griglie sono sempre visualizzate in bianco. Sono di colore grigio solo se anche lo sfondo del grafico è bianco.

Utilizzo del grafico multiplot

I grafici ed i multiplot sono essenzialmente grafici di X rispetto a Y . Ad esempio, se si sta analizzando una potenziale frode nelle richieste di concessioni agricole (come illustrato in *fraud.str* nella cartella *Demos* dell'installazione IBM SPSS Modeler), è possibile che si desideri tracciare il reddito dichiarato nella richiesta rispetto a quello stimato da una rete neurale. Una sovrapposizione, ad esempio il tipo di raccolto, consentirà di vedere se esiste una relazione tra richieste (valore o numero) e tipo di raccolto.

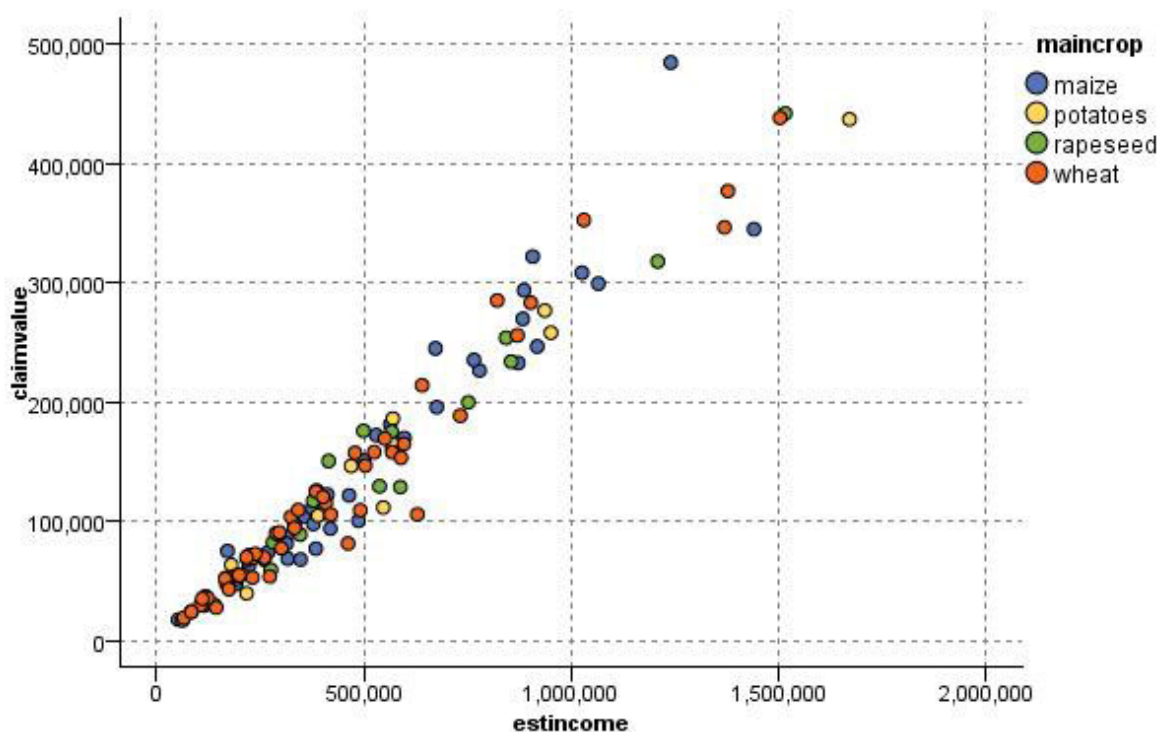


Figura 34. Grafico della relazione tra reddito stimato e valore della richiesta con il tipo raccolto principale come sovrapposizione

Dal momento che i grafici prodotti dai nodi Grafico, Multiplot e Valutazione sono visualizzazioni bidimensionali di Y rispetto a X , è facile interagire con essi selezionando le aree, contrassegnando elementi o persino disegnando sezioni mediante il mouse. È inoltre possibile generare nodi per i dati rappresentati da queste aree, sezioni o elementi. Per ulteriori informazioni, consultare l'argomento "Esplorazione dei grafici" a pagina 243.

Nodo Web

I nodi Web mostrano l'intensità delle relazioni tra valori di due o più campi simbolici. Il grafico mostra le connessioni utilizzando tipi di linea diversi a seconda dell'intensità delle connessioni. Un nodo Web può essere utilizzato, ad esempio, per analizzare la relazione tra l'acquisto di vari oggetti in un sito di e-commerce o in un tradizionale punto vendita al dettaglio.

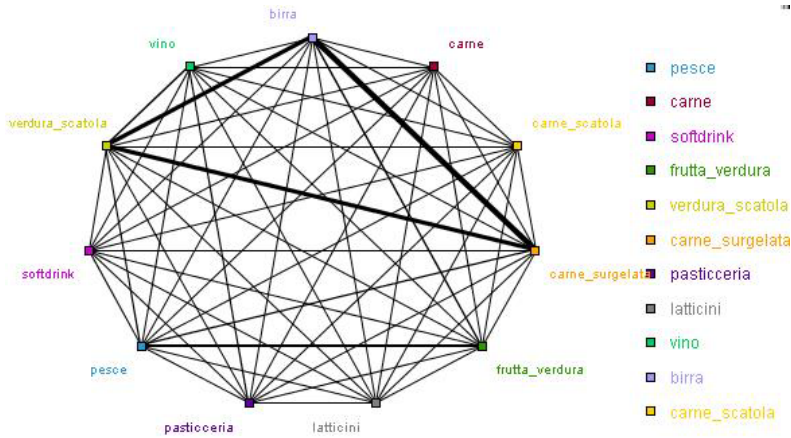


Figura 35. Grafico Web che mostra le relazioni tra l'acquisto di generi alimentari

Grafici Web diretti

I nodi Web diretti mostrano, come i nodi Web, l'intensità delle relazioni tra campi simbolici. Tuttavia, i grafici Web diretti mostrano esclusivamente le connessioni da uno o più campi Da a un singolo campo A. Le connessioni sono unidirezionali, ovvero procedono solo in un senso.

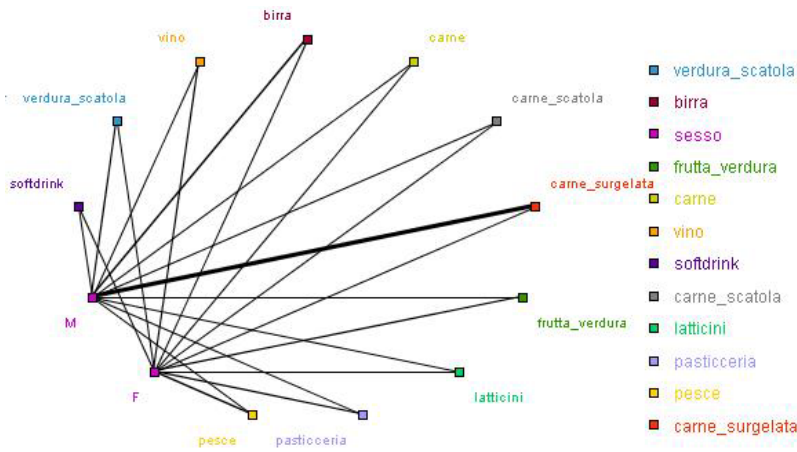


Figura 36. Grafico Web diretto che mostra la relazione tra l'acquisto di generi alimentari ed il sesso degli acquirenti

Analogamente ai nodi Web, il grafico visualizza le connessioni utilizzando tipi di linea diversi a seconda dell'intensità della connessione. Un nodo Web diretto può essere utilizzato, ad esempio, per analizzare la relazione tra il sesso dei clienti e la propensione per alcuni prodotti.

Scheda Plot del nodo Web

Web. Selezionare questa opzione per creare un grafico Web che illustri l'intensità delle relazioni tra tutti i campi specificati.

Web diretto. Selezionare questa opzione per creare un grafico Web direzionale che illustri l'intensità delle relazioni tra più campi e i valori di un campo, ad esempio il sesso o la religione. Quando questa opzione è selezionata, viene attivato un Campo A e, per maggiore chiarezza, il controllo Campi viene rinominato Campi Da.

Campo A (solo Web diretto). Selezionare un campo flag o nominale da utilizzare per un Web diretto. Verranno elencati solo i campi non esplicitamente impostati come numerici.

Campi/Campi Da. Selezionare i campi per la creazione di un grafico Web. Verranno elencati solo i campi non esplicitamente impostati come numerici. Utilizzare il pulsante Selettore di campo per selezionare più campi o per selezionarli in base al tipo.

Nota: per un Web diretto, questo comando è utilizzato per selezionare i campi Da.

Mostra solo flag veri. Selezionare questa opzione per visualizzare solo i flag veri per un campo flag. Questa opzione semplifica la visualizzazione Web e spesso viene utilizzata per i dati in cui l'occorrenza di valori positivi è molto importante.

Valori di linea. Selezionare un tipo di soglia dall'elenco a discesa.

- **Assoluto** imposta le soglie in base al numero di record con ciascuna coppia di valori.
- **Percentuali globali** mostra il numero assoluto di casi rappresentati dal collegamento come percentuale di tutte le occorrenze di ogni coppia di valori rappresentata nel grafico Web.
- **Percentuali del campo/valore più piccolo** e **Percentuali del campo/valore più grande** indicano il campo/valore da utilizzare per la valutazione delle percentuali. Ad esempio, si supponga che 100 record includano il valore *curaY* per il campo *Cura* e solo 10 abbiano il valore *BASSO* per il campo *Pressione*. Se sette record includono entrambi i valori *curaY* e *BASSO*, questa percentuale è 70% o 7% a seconda del campo a cui si fa riferimento, più piccolo (*Pressione*) o più grande (*Cura*).

Nota: per i grafici Web diretti non sono disponibili la terza e la quarta opzione sopra riportate. Sono invece disponibili le opzioni **Percentuali di campo/valore "A"** e **Percentuali di campo/valore "Da"**.

Peso maggiore per collegamenti forti. Questa opzione, selezionata per default, rappresenta la modalità comunemente utilizzata per visualizzare i collegamenti tra i campi.

Peso maggiore per collegamenti deboli. Selezionare questa opzione per invertire il significato dei collegamenti rappresentati da linee in grassetto. Questa opzione viene in genere utilizzata per il rilevamento di comportamenti fraudolenti o l'analisi di valori anomali.

Scheda Opzioni del nodo Web

La scheda Opzioni per i nodi Web contiene alcune opzioni aggiuntive per la personalizzazione del grafico di output.

Numero di collegamenti. Le seguenti opzioni vengono utilizzate per controllare il numero di collegamenti visualizzati nel grafico di output. Alcune di queste opzioni, ad esempio **Collegamenti deboli al di sotto di** e **Collegamenti forti al di sopra di**, sono disponibili anche nella finestra del grafico di output. Nel grafico finale è inoltre possibile utilizzare un dispositivo di scorrimento per modificare il numero di collegamenti visualizzati.

- **Numero massimo di collegamenti da visualizzare.** Specificare un numero che indica il numero massimo di collegamenti da mostrare nel grafico di output. Utilizzare le frecce per modificare il valore.
- **Mostra solo collegamenti al di sopra di.** Specificare un numero che indica il valore minimo per cui mostrare una connessione nel Web. Utilizzare le frecce per modificare il valore.
- **Mostra tutti i collegamenti.** Specificare per visualizzare tutti i collegamenti indipendentemente dai valori minimi o massimi. Se i campi sono numerosi, la selezione di questa opzione potrebbe determinare un aumento nel tempo di elaborazione.

Scarta se pochissimi record. Selezionare questa opzione per ignorare le connessioni supportate da un numero limitato di record. Impostare la soglia per questa opzione immettendo un numero in **Numero min. di record/righe**.

Scarta se moltissimi record. Selezionare questa opzione per ignorare le connessioni supportate da un numero di record eccessivo. Immettere un numero in **Numero max. di record/righe**.

Collegamenti deboli al di sotto di. Specificare un numero che indichi la soglia di separazione tra le connessioni deboli (linee tratteggiate) e le connessioni normali (linee normali). Tutte le connessioni inferiori a questo valore verranno considerate deboli.

Collegamenti forti al di sopra di. Specificare una soglia per le connessioni forti (linee spesse) e le connessioni regolari (linee normali). Tutte le connessioni superiori a questo valore verranno considerate forti.

Dimensione collegamento. Specificare le opzioni per il controllo della dimensione dei collegamenti:

- **La dimensione del collegamento varia continuamente.** Selezionare per visualizzare un intervallo di dimensioni del collegamento che riflette la variazione nell'intensità della connessione in base ai valori di dati effettivi.
- **La dimensione del collegamento indica le categorie forte/normale/debole.** Selezionare per visualizzare tre intensità di connessione: forte, normale e debole. I punti di interruzione per queste categorie possono essere specificati sia in questa finestra sia nel grafico finale.

Visualizzazione Web. Selezionare un tipo di visualizzazione Web:

- **Layout circolare.** Selezionare per utilizzare la visualizzazione Web standard.
- **Layout reticolare.** Selezionare per utilizzare un algoritmo per raggruppare i collegamenti più forti. Questa opzione è stata progettata per evidenziare i collegamenti forti utilizzando sia la differenziazione spaziale sia le linee ponderate.
- **Layout diretto.** Selezionare questa opzione per creare una visualizzazione Web diretta che utilizza la selezione **Campo A** della scheda Plot come direzione.
- **Layout griglia.** Selezionare questa opzione per creare una visualizzazione Web strutturata in uno schema a griglia con spaziature uniformi.

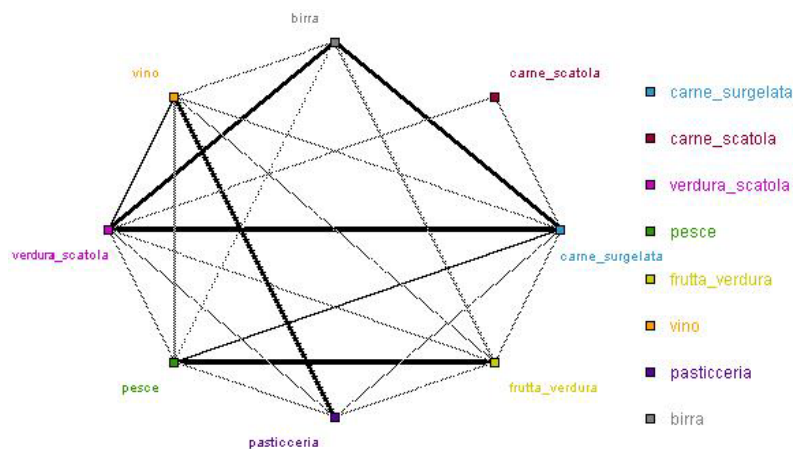


Figura 37. Gráfico Web que muestra las conexiones fuertes entre diversos tipos de alimentos

Scheda Aspetto del nodo Web

Prima de crear un gráfico es posible especificar las opciones de l'aspetto.

Titolo. Immettere il testo da utilizzare come titolo del gráfico.

Sottotitolo. Immettere il testo da utilizzare come sottotitolo del gráfico.

Didascalia. Immettere il testo che verrà utilizzato per la didascalia del gráfico.

Mostra legenda. Specifica se la legenda viene visualizzata. Per i grafici con un numero elevato di campi, può essere consigliabile nascondere la legenda per migliorare l'aspetto del grafico.

Utilizza etichette come nodi. Include il testo delle etichette in ogni nodo anziché visualizzare le etichette adiacenti. Nei grafici con un numero limitato di campi, questa opzione può aumentare la leggibilità del grafico.



Figura 38. Grafico Web che visualizza le etichette come nodi

Utilizzo di un grafico Web

I nodi Web sono utilizzati per mostrare l'intensità delle relazioni tra i valori di due o più campi simbolici. Le connessioni sono visualizzate in un grafico con tipi di linea diversi a seconda dell'intensità. Un nodo Web può essere utilizzato, ad esempio, per analizzare la relazione tra i livelli di colesterolo, la pressione sanguigna e la cura che si è rivelata efficace nel trattamento della malattia del paziente.

- Le connessioni forti sono rappresentate da una linea spessa. Questo tipo di linea indica che i due valori sono fortemente correlati e devono essere sottoposti a un'analisi più approfondita.
- Le connessioni medie sono rappresentate da una linea di spessore normale.
- Le connessioni deboli sono rappresentate da una linea tratteggiata.
- L'assenza di una linea di collegamento tra due valori indica che i due valori non ricorrono mai nello stesso record oppure sono contemporaneamente presenti in un numero di record inferiore alla soglia specificata nella finestra di dialogo del nodo Web.

Dopo aver creato un nodo Web, è possibile utilizzare diverse opzioni per modificare la visualizzazione del grafico e generare nodi per ulteriori analisi.

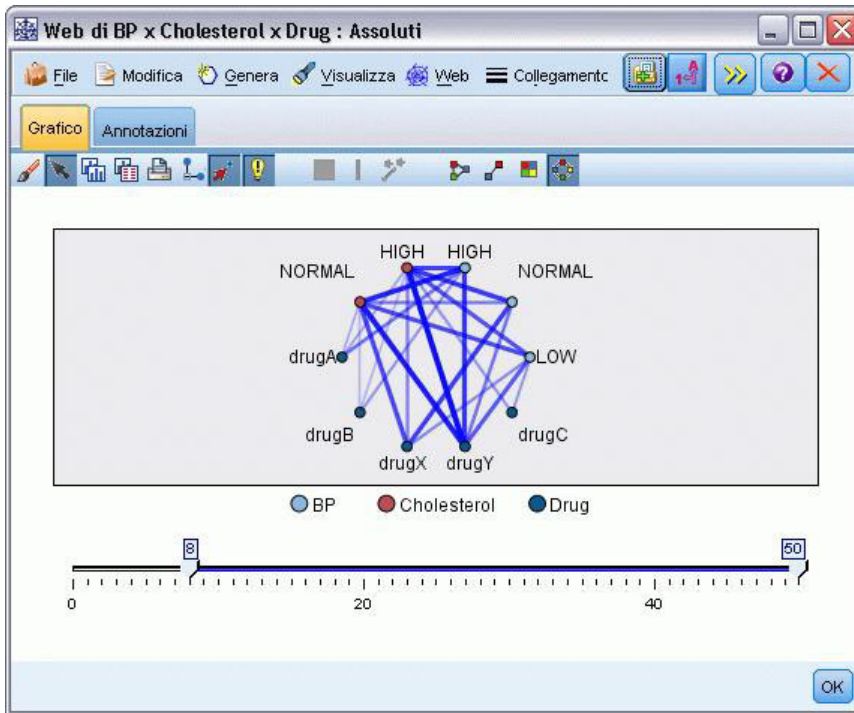


Figura 39. Grafico Web che indica un numero di relazioni forti, ad esempio una pressione sanguigna normale con curaX e un livello di colesterolo elevato con curaY

Sia per i nodi Web che per i nodi Web diretto è possibile:

- Modificare il layout della visualizzazione Web.
- Nascondere punti per semplificare la visualizzazione.
- Modificare le soglie che controllano gli stili di linea.
- Evidenziare le linee tra i valori per indicare una relazione "selezionata".
- Generare un nodo Seleziona per uno o più record "selezionati" o un nodo Ricava flag associato a una o più relazioni nel grafico Web.

Per modificare i punti

- **Spostare** i punti facendo clic con il pulsante del mouse su un punto e trascinandolo in una nuova posizione. Il grafico Web verrà ridisegnato per riflettere la nuova posizione.
- **Nascondere** i punti facendo clic con il pulsante destro del mouse su un punto nel grafico Web e scegliendo **Nascondi** o **Nascondi e ripianifica** dal menu di scelta rapida. **Nascondi** nasconde semplicemente il punto selezionato e tutte le linee ad esso associate. **Nascondi e ripianifica** ridisegna il grafico Web in base alle modifiche apportate. Qualsiasi spostamento manuale viene annullato.
- **Mostrare** tutti i punti nascosti scegliendo **Mostra tutto** o **Mostra tutto e ripianifica** dal menu Web nella finestra del grafico. Se si seleziona **Mostra tutto e ripianifica**, il grafico Web viene ridisegnato per includere tutti i punti nascosti in precedenza e le relative connessioni.

Per selezionare o evidenziare le linee

Le linee selezionate sono evidenziate in rosso.

1. Per selezionare una sola linea, fare clic su di essa con il pulsante sinistro del mouse.
2. Per selezionare più linee, effettuare una delle seguenti operazioni:
 - Utilizzando il cursore, disegnare un cerchio attorno ai punti di cui si desidera selezionare le linee.
 - Tenere premuto il tasto Ctrl e fare clic con il pulsante sinistro del mouse sulle singole linee da selezionare.

Per deselezionare tutte le linee selezionate, fare clic sullo sfondo del grafico oppure scegliere **Cancella selezione** dal menu Web nella finestra del grafico.

Per visualizzare il grafico Web utilizzando un layout diverso

Dal menu Web, scegliere **Layout circolare**, **Layout reticolare**, **Layout diretto** o **Layout griglia** per modificare il layout del grafico.

Per attivare e disattivare il dispositivo di scorrimento dei collegamenti

Dal menu Visualizza, scegliere **Dispositivo di scorrimento dei collegamenti**.

Per selezionare o generare un flag per i record di una singola relazione

1. Fare clic con il pulsante destro del mouse sulla linea che rappresenta la relazione desiderata.
2. Dal menu di scelta rapida scegliere **Genera nodo Seleziona per il collegamento** o **Genera nodo Ricava per il collegamento**.

Nell'area del flusso viene aggiunto automaticamente un nodo Seleziona o un nodo Ricava con le opzioni appropriate e le condizioni specificate:

- Il nodo Seleziona seleziona tutti i record nella relazione selezionata.
- Il nodo Ricava genera un flag che indica se la relazione selezionata è vera per i record dell'intero insieme di dati. Il nome del campo flag è composto dai due valori presenti nella relazione uniti da un carattere di sottolineatura, ad esempio *BASSO_curaC* o *curaC_BASSO*.

Per selezionare o generare un flag per i record di un gruppo di relazioni

1. Selezionare la linea o le linee nella visualizzazione Web che rappresentano le relazioni selezionate.
 2. Dal menu Genera nella finestra del grafico, scegliere **Nodo Seleziona ("E")**, **Nodo Seleziona ("O")**, **Nodo Ricava ("E")** o **Nodo Ricava ("O")**.
- I nodi "O" sono basati sulla separazione delle condizioni. Ciò significa che il nodo verrà applicato ai record per i quali vale una qualsiasi delle relazioni selezionate.
 - I nodi "E" sono basati sull'unione delle condizioni. Ciò significa che il nodo verrà applicato solo ai record per i quali valgono tutte le relazioni selezionate. Se le relazioni selezionate si escludono a vicenda, viene generato un errore.

Al termine della selezione, nell'area del flusso viene automaticamente aggiunto un nodo Seleziona o Ricava, con le opzioni appropriate e le condizioni specificate.

Modifica delle soglie di un grafico Web

Al termine della creazione di un grafico Web, è possibile modificare le soglie che controllano gli stili della linea utilizzando il dispositivo di scorrimento della barra degli strumenti per modificare la linea minima visibile. È anche possibile visualizzare opzioni aggiuntive di soglia facendo clic sul pulsante con due frecce gialle disponibile nella barra degli strumenti per espandere la finestra del grafico Web. Quindi fare clic sulla scheda **Controlli** per visualizzare le opzioni aggiuntive.

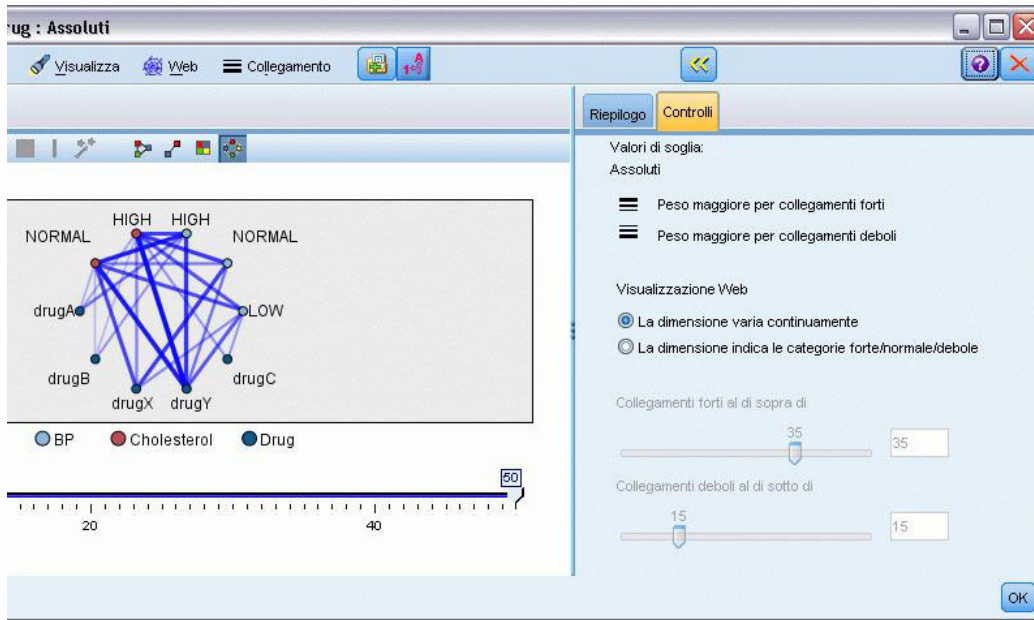


Figura 40. Finestra espansa con le opzioni di visualizzazione e di soglia

Valori di soglia. Mostra il tipo di soglia selezionato durante la creazione nella finestra di dialogo del nodo Web.

Peso maggiore per collegamenti forti. Selezionato per default, questo è il metodo standard per la visualizzazione dei collegamenti tra i campi.

Peso maggiore per collegamenti deboli. Selezionare per invertire il significato dei collegamenti visualizzati in grassetto. Questa opzione viene in genere utilizzata per il rilevamento di comportamenti fraudolenti o l'analisi di valori anomali.

Visualizzazione Web. Specificare le opzioni per il controllo della dimensione dei collegamenti nel grafico di output:

- **La dimensione varia continuamente.** Selezionare per visualizzare un intervallo di dimensioni del collegamento che riflette la variazione nell'intensità della connessione in base ai valori di dati effettivi.
- **La dimensione indica le categorie forte/normale/debole.** Selezionare per visualizzare tre intensità di connessione: forte, normale e debole. I punti di interruzione per queste categorie possono essere specificati sia in questa finestra sia nel grafico finale.

Collegamenti forti al di sopra di. Specificare una soglia per le connessioni forti (linee spesse) e le connessioni regolari (linee normali). Tutte le connessioni superiori a questo valore verranno considerate forti. Utilizzare il dispositivo di scorrimento per modificare il valore o immettere un numero nel campo.

Collegamenti deboli al di sotto di. Specificare un numero che indichi la soglia di separazione tra le connessioni deboli (linee tratteggiate) e le connessioni normali (linee normali). Tutte le connessioni inferiori a questo valore verranno considerate deboli. Utilizzare il dispositivo di scorrimento per modificare il valore o immettere un numero nel campo.

Dopo aver modificato le soglie per un grafico Web, è possibile ripianificare o ridisegnare la visualizzazione Web con i nuovi valori di soglia tramite il menu Web posto nella barra degli strumenti del grafico Web. Dopo aver individuato le impostazioni che rivelano gli schemi più significativi, è possibile aggiornare le impostazioni originali nel nodo Web (denominato anche nodo Web padre) scegliendo **Aggiorna nodo padre** dal menu Web nella finestra del grafico.

Creazione di un riepilogo Web

È possibile creare un documento di riepilogo Web in cui siano elencati i collegamenti forti, medi e deboli, facendo clic sul pulsante con due frecce gialle nella barra degli strumenti per espandere la finestra del grafico Web. Quindi fare clic sulla scheda **Riepilogo** per visualizzare le tabelle per ogni tipo di collegamento. Le tabelle possono essere espanse e compresse utilizzando i relativi pulsanti con le frecce.

Per stampare i riepilogo, scegliere le seguenti opzioni dal menu nella finestra del grafico Web:

File > Stampa riepilogo

Nodo del grafico temporale

I nodi del grafico temporale consentono di visualizzare una o più serie temporali rappresentate nel tempo. Le serie rappresentate devono contenere valori numerici e si presuppone che ricorrano in un intervallo di tempo in cui i periodi sono uniformi. Il nodo Intervalli di tempo viene utilizzato in genere prima del nodo del grafico temporale per creare un campo *EtichettaTempo*, utilizzato per default per identificare l'asse delle x nei grafici. Per ulteriori informazioni, consultare l'argomento "Nodo Intervalli di tempo" a pagina 151.

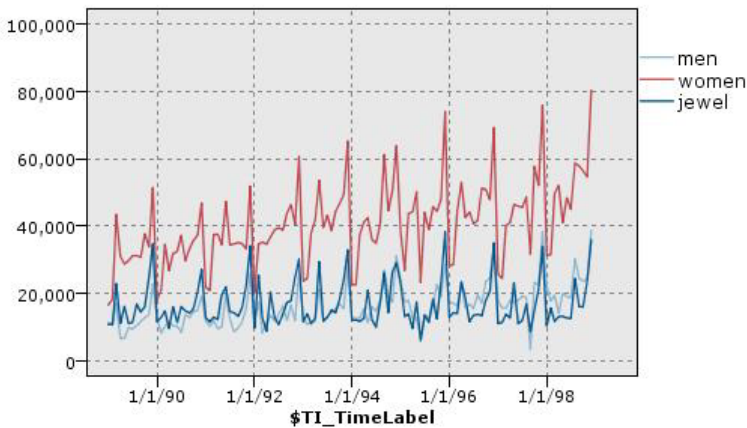


Figura 41. Rappresentazione grafica delle vendite di abbigliamento e gioielli maschili e femminili nel tempo

Creazione di interventi ed eventi

È possibile creare campi di evento e di intervento dal Plot tempo generando un nodo Ricava (flag o nominale) dai menu di scelta rapida. Per esempio, si può creare un campo di evento per uno sciopero ferroviario, dove lo stato guida è Vero se l'evento si verifica e Falso in caso contrario. Nel caso di un campo di Intervento, per esempio per un aumento di prezzo, si può invece utilizzare un nodo Crea variabile conteggio per identificare la data dell'aumento, con il valore 0 per il prezzo vecchio e il valore 1 per quello nuovo. Per ulteriori informazioni, consultare l'argomento "Nodo Ricava" a pagina 127.

Scheda Plot del nodo Plot tempo

Plot. Consente di scegliere il modo in cui rappresentare graficamente i dati di serie temporali.

- **Serie selezionata.** Rappresenta i valori della serie temporale selezionata. Se si seleziona questa opzione quando si rappresentano gli intervalli di confidenza, deselegionare la casella di controllo **Normalizza**.
- **Modelli di serie temporali selezionati.** Utilizzata con un modello di serie temporali, questa opzione rappresenta tutti i campi correlati (valori effettivi e previsti e intervalli di confidenza) per una o più

serie temporali selezionate. La selezione di questa opzione comporta la disattivazione di alcune altre opzioni nella finestra di dialogo. Si tratta dell'opzione ideale per la rappresentazione degli intervalli di confidenza.

Serie. Selezionare uno o più campi contenenti i dati di serie temporali da rappresentare nel grafico. I dati devono essere numerici.

Etichetta dell'asse X. Scegliere l'etichetta di default o un singolo campo da utilizzare come etichetta dell'asse x nelle rappresentazioni grafiche. Se si seleziona l'opzione Default, il sistema utilizza il campo EtichettaTempo creato a partire da un nodo Intervalli di tempo situato a monte o, in assenza del nodo Intervalli di tempo, da numeri interi sequenziali. Per ulteriori informazioni, consultare l'argomento "Nodo Intervalli di tempo" a pagina 151.

Visualizza serie in riquadri distinti. Consente di specificare se le singole serie devono essere visualizzate in riquadri distinti. In alternativa, se si decide di non utilizzare i riquadri, tutte le serie temporali vengono rappresentate sullo stesso grafico e le funzioni di livellamento non sono disponibili. Quando si rappresentano tutte le serie temporali sullo stesso grafico, ogni serie è contraddistinta da un colore diverso.

Normalizza. Selezionare per ridimensionare tutti i valori Y in base all'intervallo 0–1 per la visualizzazione nel grafico. La normalizzazione consente di visualizzare in modo chiaro la relazione tra linee che potrebbero risultare non completamente visibili a causa di differenze nell'intervallo dei valori di ogni serie ed è consigliata quando si rappresentano più linee nello stesso grafico o quando si confrontano plot in riquadri affiancati (la normalizzazione non è necessaria quando tutti i valori di dati rientrano in un intervallo simile).

Visualizzazione. Seleziona uno o più elementi da visualizzare nella rappresentazione grafica. È possibile scegliere tra linee, punti e livellamenti (LOESS). I livellamenti sono disponibili solo se le serie vengono visualizzate in riquadri distinti. Per default è selezionato l'elemento linea. Verificare di selezionare almeno un elemento da rappresentare prima di eseguire il nodo Grafico; in caso contrario, il sistema restituisce un errore che indica che non è stato selezionato alcun elemento da tracciare.

Limita record. Selezionare questa opzione per limitare il numero dei record da rappresentare nel grafico. Specificare il numero di record (letti dall'inizio del file di dati) che dovranno essere rappresentati nell'opzione **Numero massimo di record per il plot**. Per default, tale numero è impostato su 2000. Per rappresentare gli ultimi n record del file di dati, è possibile utilizzare un nodo Ordina prima di questo nodo per ordinare i record in ordine decrescente in base all'ora.

Scheda Aspetto del nodo Plot tempo

Prima di creare un grafico è possibile specificarne le opzioni dell'aspetto.

Titolo. Immettere il testo da utilizzare come titolo del grafico.

Sottotitolo. Immettere il testo da utilizzare come sottotitolo del grafico.

Didascalia. Immettere il testo che verrà utilizzato per la didascalia del grafico.

Etichetta X. Accettare l'etichetta dell'asse x (orizzontale) generata automaticamente oppure selezionare **Personalizzato** per specificare un'etichetta personalizzata.

Etichetta Y. Accettare l'etichetta dell'asse y (verticale) generata automaticamente oppure selezionare **Personalizzato** per specificare un'etichetta personalizzata.

Visualizza griglia. Questa opzione è selezionata per default e consente di visualizzare una griglia dietro il grafico per stabilire più facilmente i punti di interruzione di aree e sezioni. Le griglie sono sempre visualizzate in bianco. Sono di colore grigio solo se anche lo sfondo del grafico è bianco.

Layout. Solo per i Plot tempo, è possibile specificare se i valori temporali devono essere rappresentati su un asse orizzontale o verticale.

Utilizzo del grafico tempo

Dopo aver creato il grafico tempo è possibile utilizzare diverse opzioni per modificare la visualizzazione del grafico e generare nodi per ulteriori analisi. Per ulteriori informazioni, consultare l'argomento "Esplorazione dei grafici" a pagina 243.

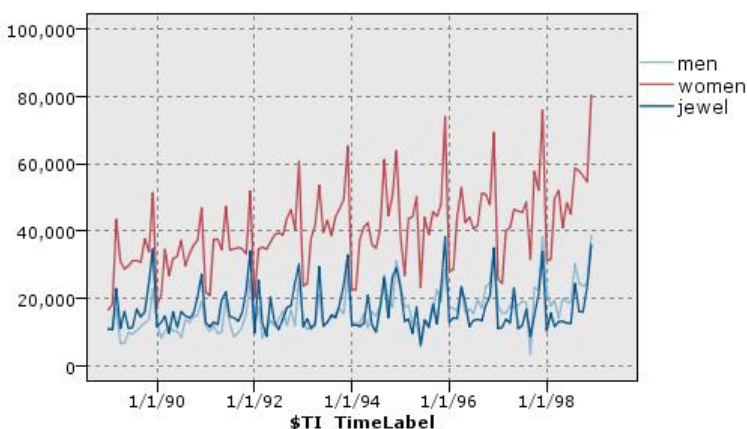


Figura 42. Rappresentazione grafica delle vendite di abbigliamento e gioielli maschili e femminili nel tempo

Dopo aver creato un Plot tempo, definito le sezioni ed esaminato i risultati, è possibile utilizzare le opzioni del menu Genera e il menu di scelta rapida per creare nodi Seleziona o Ricava. Per ulteriori informazioni, consultare l'argomento "Generazioni di nodi da grafici" a pagina 251.

Nodo Valutazione

Il nodo Valutazione consente di valutare e mettere a confronto modelli predittivi per scegliere il miglior modello per l'applicazione da utilizzare. I grafici prodotti dal nodo Valutazione (o grafici di valutazione) mostrano il comportamento dei modelli nella previsione di particolari risultati. Questi grafici ordinano i record in base al valore previsto e alla confidenza della previsione, quindi li suddividono in gruppi di uguale dimensione (**quantili**) e infine rappresentano il valore del criterio di business per ciascun quantile, dal più alto al più basso. I modelli multipli sono mostrati nel grafico come linee separate.

I risultati vengono gestiti definendo un valore specifico o un intervallo di valori come **risultato**. I risultati in genere indicano un qualche successo (ad esempio una vendita a un cliente) o un evento significativo (ad esempio una diagnosi medica specifica). Per definire i criteri dei risultati, utilizzare la scheda Opzioni della finestra di dialogo oppure utilizzare i criteri di default, come segue:

- I campi di output **flag** sono di semplice interpretazione; i risultati corrispondono ai valori *true*.
- Nei campi di output **nominali** il primo valore dell'insieme definisce un risultato.
- Nei campi di output **continui** i risultati equivalgono ai valori superiori al valore intermedio dell'intervallo del campo.

Sono disponibili sei tipi di grafici di valutazione, ognuno dei quali enfatizza un criterio di valutazione diverso.

Grafici dei profitti

I guadagni sono rappresentati dalla percentuale di risultati complessivi in ogni quantile. I guadagni sono calcolati come $(\text{numero di risultati nel quantile} / \text{numero totale di risultati}) \times 100\%$.

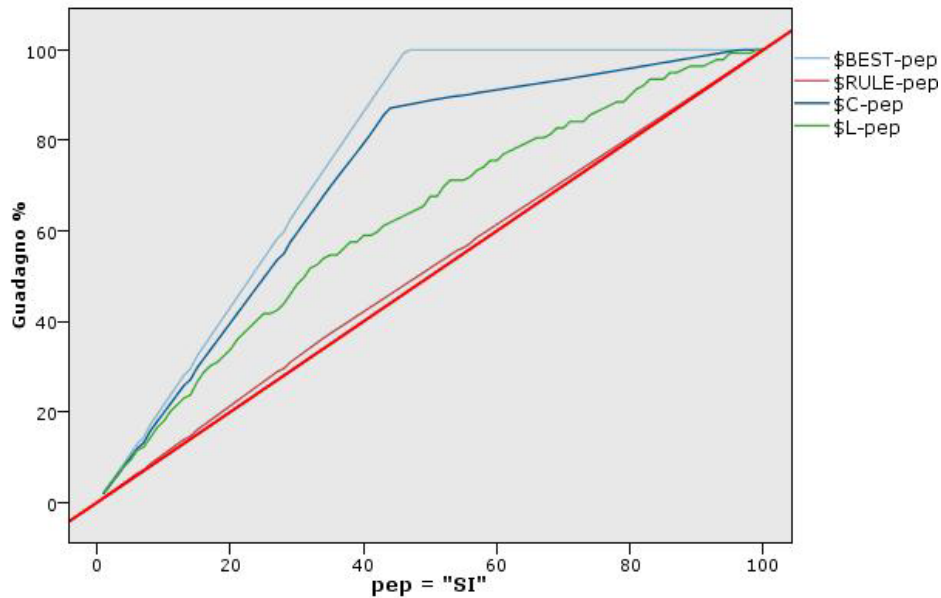


Figura 43. Grafico dei profitti (cumulato) con linea base, linea migliore e regola di business

Grafici guadagno cumulativo

Il grafico guadagno cumulativo confronta la percentuale di record che in ogni quantile corrispondono a risultati con la percentuale globale di risultati nei dati di addestramento. La risposta viene calcolata dividendo i risultati nel quantile per i record nel quantile e quindi dividendo il risultato ottenuto per il risultato della divisione dei risultati totali per i record totali.

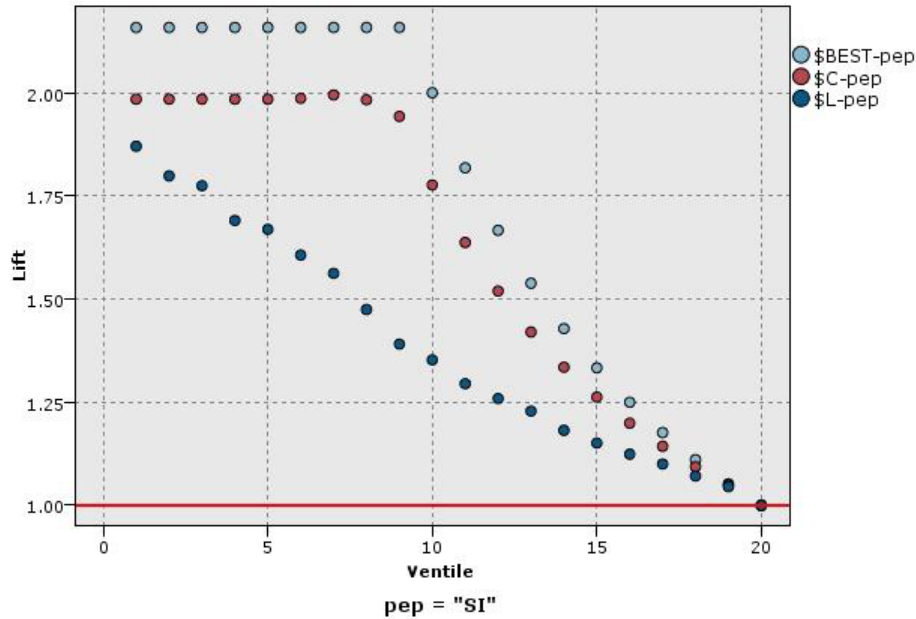


Figura 44. Grafico guadagno (cumulato) con l'utilizzo di punti e linea migliore

Grafici delle risposte

La risposta è rappresentata dalla percentuale di record nel quantile corrispondenti a risultati. La risposta è calcolata come $(\text{risultati nel quantile} / \text{record nel quantile}) \times 100\%$.

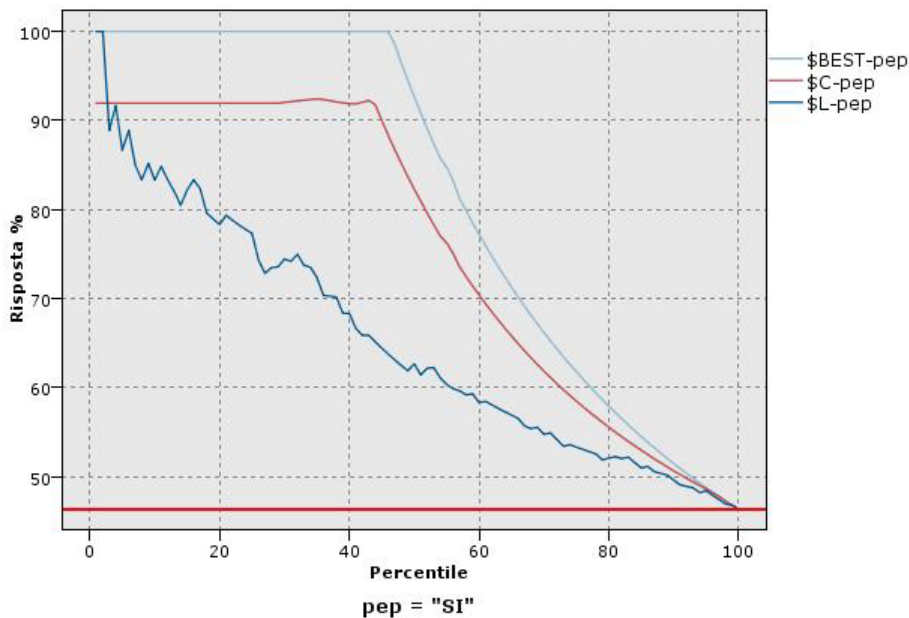


Figura 45. Grafico delle risposte (cumulato) con linea migliore

Grafici di profitto

Il profitto equivale alle **entrate** relative a ogni record meno il **costo** per il record. I profitti di un quantile sono semplicemente la somma dei profitti di tutti i record nel quantile. Si presuppone che le entrate

vengano applicate solo ai risultati, mentre i costi vengono applicati a tutti i record. Profitti e costi possono essere fissi oppure definiti da campi nei dati. I profitti sono calcolati come (somma delle entrate per i record nel quantile – somma dei costi per i record nel quantile).

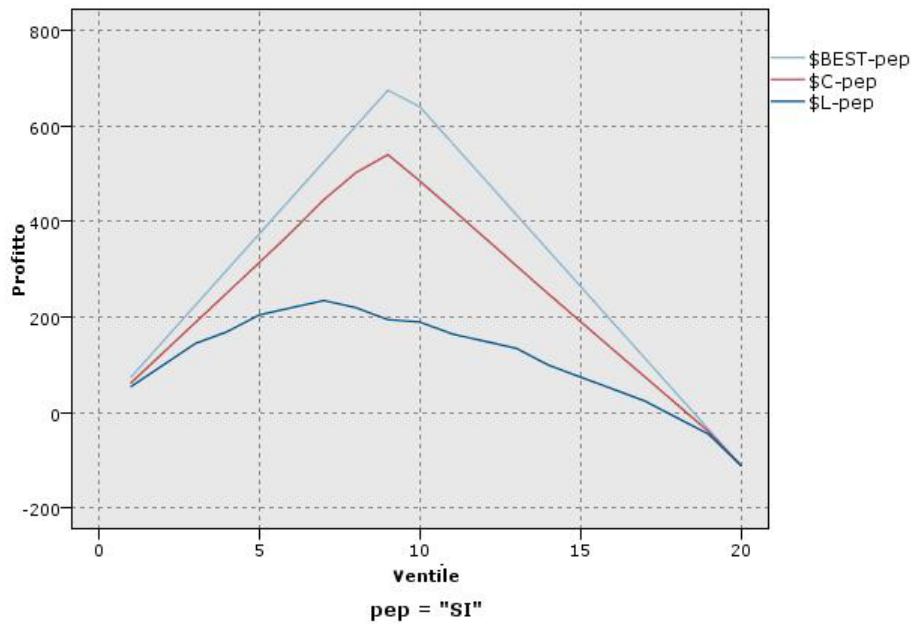


Figura 46. Grafico dei profitti (cumulato) con linea migliore

Grafici ROI

Il valore ROI (Return On Investment) è simile al profitto in quanto implica la definizione di entrate e costi. Il ROI confronta i profitti con i costi per il quantile. Il valore ROI viene calcolato come (profitti per quantile / costi per quantile) \times 100%.

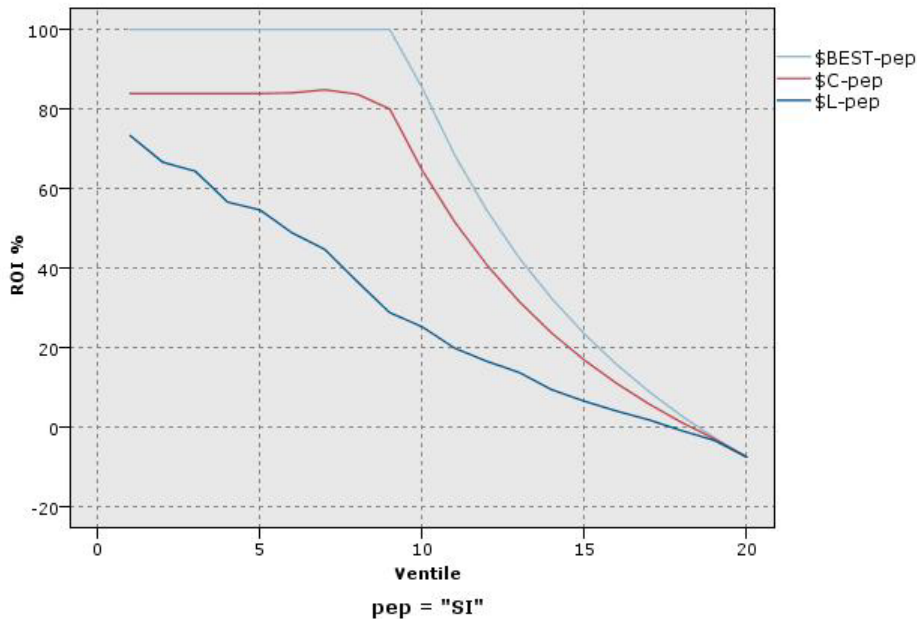


Figura 47. Grafico ROI (cumulato) con linea migliore

Grafici ROC

ROC (receiver operator characteristic) può essere utilizzato solo con classificatori binari. ROC può essere utilizzato per visualizzare, organizzare e selezionare i classificatori in base alle relative prestazioni. Un grafico ROC traccia graficamente il tasso (o la sensibilità) di veri positivi a fronte del tasso di falsi positivi del classificatore. Un grafico ROC rappresenta i compromessi relativi tra i vantaggi (veri positivi) ed i costi (falsi positivi). Un vero positivo è un'istanza che rappresenta un risultato e viene classificata come risultato. Pertanto il tasso dei veri positivi viene calcolato come numero di veri positivi/numero di istanze che sono effettivamente risultati. Un falso positivo è un'istanza che rappresenta un risultato mancato e viene classificata come risultato. Pertanto il tasso dei falsi positivi viene calcolato come numero di falsi positivi/numero di istanze che sono effettivamente mancati risultati.

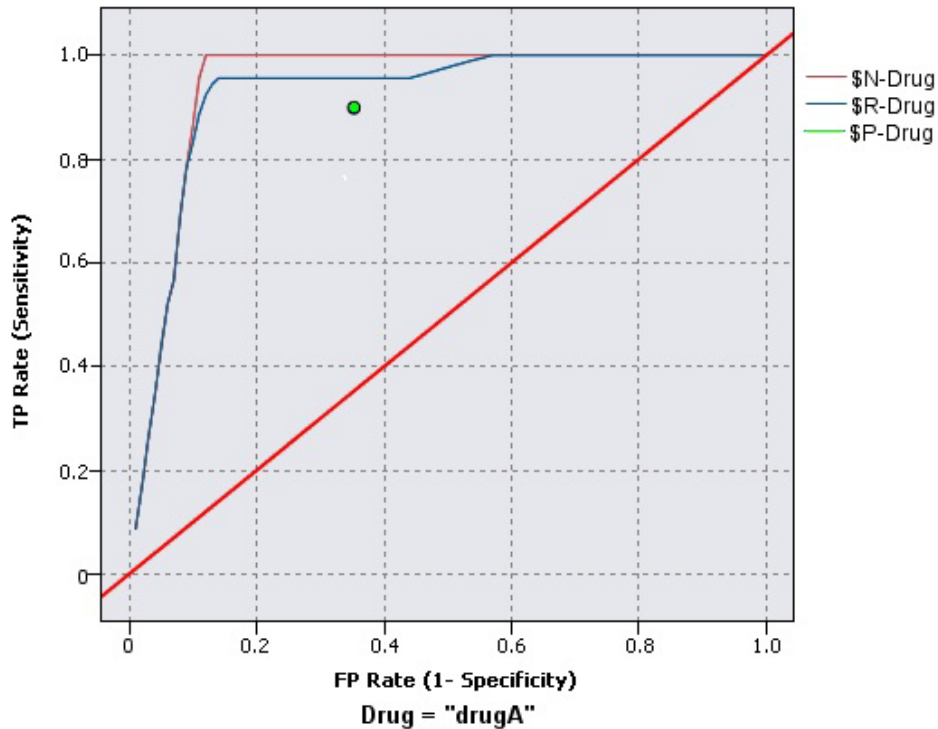


Figura 48. Grafico ROC con linea migliore

I grafici di valutazione possono essere anche cumulati, per cui ogni punto equivale al valore del quantile corrispondente più tutti i quantili superiori. I grafici cumulativi in genere offrono una rappresentazione migliore delle performance globali dei modelli, mentre i grafici non cumulativi risultano spesso più efficaci per mettere in evidenza particolari aree problematiche dei modelli.

Scheda Plot del nodo Valutazione

Tipo di grafico. Selezionare uno dei seguenti tipi: **Guadagni**, **Risposta**, **Guadagno cumulativo**, **Profitto**, **ROI** (return on investment) o **ROC** (receiver operator characteristic).

Grafico cumulato. Selezionare questa opzione per creare un grafico cumulativo. I valori presenti nei grafici cumulativi vengono rappresentati per ogni quantile più tutti i quantili superiori. (**Grafico cumulato** non è disponibile per i grafici ROC).

Includi linea base. Selezionare questa opzione per includere una linea base nel grafico, che indica una distribuzione casuale dei risultati quando la confidenza diventa irrilevante (**Includi baseline** non è disponibile per i grafici Profitto e ROI).

Includi linea migliore. Selezionare questa opzione per includere una linea migliore nel grafico, che indica una confidenza perfetta (i risultati sono pari al 100% dei casi). (**Includi linea migliore**) non è disponibile per i grafici ROC).

Utilizza criteri profitto per tutti i tipi di grafici. Selezionare questa opzione per utilizzare i criteri di profitto (costo, entrate e peso) quando si calcolano le misure di valutazione, invece del normale conteggio dei risultati. Per i modelli con determinati obiettivi numerici come ad esempio un modello che prevede le entrate ottenute da un cliente in risposta ad un'offerta, il valore del campo obiettivo fornisce una misurazione migliore delle prestazioni del modello rispetto al calcolo dei risultati. La selezione di questa opzione abilita i campi **Costi**, **Entrate** e **Peso** per i grafici Guadagni, Risposta e Guadagno cumulativo. Per utilizzare i criteri profitto per questi tre tipi di grafici si consiglia di impostare **Entrate** come campo

obiettivo, **Costi** su 0.0 in modo che il profitto sia uguale alle entrate e specificare una condizione di risultati definita dall'utente su "true" così che tutti i record vengono calcolati come risultati. (**Utilizza criteri profitto per tutti i tipi di grafici** non è disponibile per i grafici ROC).

Trova campi previsti/predittore mediante. Selezionare **Metadati del campo di output del modello** per cercare i campi previsti nel grafico utilizzando i relativi metadati o selezionare **Formato del nome campo** per eseguire la ricerca in base ai nomi.

Campi Punteggio plot. Selezionare questa casella di controllo per attivare il selettore di campi punteggio. Selezionare quindi uno o più campi punteggio intervallo o continui, ovvero campi che non sono modelli predittivi in senso stretto ma che potrebbero essere utili per classificare i record a seconda della loro propensione a essere risultati. Il nodo Valutazione può confrontare qualunque combinazione di uno o più campi punteggio con uno o più modelli predittivi. Un esempio tipico potrebbe essere il confronto di vari campi RFM con il modello predittivo migliore.

Obiettivo. Selezionare il campo obiettivo mediante il selettore di campo. Scegliere tutti i campi flag o nominali istanziati con due o più valori.

Nota: questo campo obiettivo è valido solo per i campi punteggio (i modelli predittivi definiscono da soli i propri obiettivi) e viene ignorato se nella scheda Opzioni viene impostato un criterio personalizzato per i risultati.

Suddividi per partizione. Se si utilizza un campo partizione per suddividere i record in campioni di addestramento, di test e di convalida, selezionare questa opzione per visualizzare un grafico di valutazione distinto per ogni partizione. Per ulteriori informazioni, consultare l'argomento "Nodo Partizione" a pagina 146.

Nota: se si suddivide per partizione, i record con valori null nel campo partizione saranno esclusi dalla valutazione. Ciò non rappresenta un problema se si utilizza un nodo Partizione, poiché i nodi Partizione non generano valori null.

Grafico. Selezionare dall'elenco a discesa la dimensione dei quantili da rappresentare nel grafico. Le opzioni includono **Quartili, Quintili, Decili, Ventili, Percentili** e **1000-tili**. (**Grafico** non è disponibile per i grafici ROC).

Stile. Selezionare **Linea** o **Punto**.

Per tutti i tipi di grafici tranne i grafici ROC, sono disponibili controlli aggiuntivi che consentono di specificare costi, entrate e pesi.

- **Costi.** Specificare il costo associato a ciascun record. È possibile selezionare un costo **Fisso** o **Variabile**. Per i costi fissi, specificare il valore del costo. Per i costi variabili, fare clic sul pulsante Selettore di campo per selezionare un campo come campo dei costi. (**Costi** non è disponibile per i grafici ROC).
- **Entrate.** Specificare le entrate associate a ogni record che rappresenta un risultato. È possibile selezionare un costo **Fisso** o **Variabile**. Per entrate fisse, specificare il valore delle entrate. Per entrate variabili, fare clic sul pulsante Selettore di campo per selezionare un campo come campo delle entrate. (**Entrate** non è disponibile per i grafici ROC).
- **Peso.** Se i record nei dati rappresentano più di una unità, è possibile utilizzare i pesi per modificare i risultati. Specificare il peso associato a ciascun record, utilizzando i pesi **Fisso** o **Variabile**. Per pesi fissi, specificare il valore del peso (numero di unità per record). Per pesi variabili, fare clic sul pulsante Selettore di campo per selezionare un campo come campo dei pesi. (**Peso** non è disponibile per i grafici ROC).

Scheda Opzioni del nodo Valutazione

La scheda Opzioni per i grafici di valutazione garantisce la flessibilità nella definizione dei risultati, dei criteri di calcolo del punteggio e delle regole di business visualizzati nel grafico. È anche possibile impostare opzioni aggiuntive per l'esportazione dei risultati della valutazione del modello.

Risultato definito dall'utente. Selezionare questa opzione per specificare una condizione personalizzata da utilizzare per indicare un risultato. Questa opzione è utile per definire il risultato desiderato anziché dedurlo dal tipo di campo obiettivo e dall'ordine dei valori.

- **Condizione.** Quando è stato selezionato **Risultato definito dall'utente**, è necessario specificare un'espressione CLEM per una condizione di risultato. Per esempio, @TARGET = "S \hat{I} " è una condizione valida che indica che un valore S \hat{I} per il campo obiettivo verrà considerato come un risultato nella valutazione. La condizione specificata verrà utilizzata per tutti i campi obiettivo. Per creare una condizione, digitare direttamente nel campo o utilizzare il Generatore di espressioni per generare un'espressione di condizione. Se i dati sono istanziati, è possibile immettere valori direttamente dal Generatore espressioni.

Punteggio definito dall'utente. Selezionare questa opzione per specificare una condizione utilizzata per calcolare il punteggio dei casi prima di assegnarli ai quantili. Il punteggio di default viene calcolato in base al valore previsto e alla confidenza. Utilizzare il campo Espressione per creare un'espressione di punteggio personalizzata.

- **Espressione.** Specificare un'espressione CLEM utilizzata per il calcolo del punteggio. Ad esempio, se un output numerico nell'intervallo 0–1 viene ordinato in modo che i valori più bassi sono migliori di quelli più alti, è possibile definire un risultato come @TARGET < 0.5 ed il punteggio associato come 1 - @PREDICTED. L'espressione del punteggio deve generare un valore numerico. Per creare una condizione, digitare direttamente nel campo o utilizzare il Generatore di espressioni per generare un'espressione di condizione.

Includi regola di business. Selezionare questa opzione per specificare una condizione di regola che rifletta i criteri desiderati. Per esempio, è possibile visualizzare una regola per tutti i casi in cui mutuo = "Y" e reddito >= 33000. Le regole di business vengono tracciate sul grafico ed etichettate nella chiave come *Regola*. (**Includi regola di business** non è supportato per i grafici ROC).

- **Condizione.** Specificare un'espressione CLEM utilizzata per definire una regola di business nel grafico di output. Digitare direttamente nel campo o utilizzare il Generatore di espressioni per generare un'espressione di condizione. Se i dati sono istanziati, è possibile immettere valori direttamente dal Generatore espressioni.

Esporta risultati nel file. Selezionare questa opzione per esportare i risultati della valutazione del modello in un file di testo delimitato. È possibile leggere questo file per eseguire analisi specializzate sui valori calcolati. Impostare le seguenti opzioni per l'esportazione:

- **Nome file.** Immettere il nome del file di output. Utilizzare il pulsante con i puntini di sospensione (...) per passare alla cartella desiderata.
- **Delimitatore.** Immettere un carattere, ad esempio una virgola o uno spazio, da utilizzare come separatore di campo.

Includi nomi campi. Selezionare questa opzione per includere i nomi dei campi come prima riga del file di output.

Nuova riga dopo ogni record. Selezionare questa opzione se si desidera che ogni record inizi in una nuova riga.

Scheda Aspetto del nodo Valutazione

Prima di creare un grafico è possibile specificarne le opzioni dell'aspetto.

Titolo. Immettere il testo da utilizzare come titolo del grafico.

Sottotitolo. Immettere il testo da utilizzare come sottotitolo del grafico.

Testo. Accettare l'etichetta di testo generata automaticamente o selezionare **Personalizzato** per specificare un'etichetta.

Etichetta X. Accettare l'etichetta dell'asse x (orizzontale) generata automaticamente oppure selezionare **Personalizzato** per specificare un'etichetta personalizzata.

Etichetta Y. Accettare l'etichetta dell'asse y (verticale) generata automaticamente oppure selezionare **Personalizzato** per specificare un'etichetta personalizzata.

Visualizza griglia. Questa opzione è selezionata per default e consente di visualizzare una griglia dietro il grafico per stabilire più facilmente i punti di interruzione di aree e sezioni. Le griglie sono sempre visualizzate in bianco. Sono di colore grigio solo se anche lo sfondo del grafico è bianco.

Lettura dei risultati di un grafico di valutazione relativo a un modello

L'interpretazione di un grafico di valutazione dipende entro certi limiti dal tipo di grafico, tuttavia alcune caratteristiche sono comuni a tutti i grafici di valutazione. Nel caso dei grafici cumulativi, le linee superiori indicano i modelli migliori, in particolare nella parte sinistra del grafico. In molti casi, quando si confrontano più modelli le linee si incrociano, pertanto un modello risulterà più alto in una parte del grafico e un altro risulterà più alto in un'altra parte del grafico. In questi casi è necessario definire la parte del campione desiderata (la parte che individua un punto sull'asse x) quando si sceglie il modello da utilizzare.

La maggior parte dei grafici non cumulativi sarà molto simile. Per modelli efficaci, i grafici non cumulativi devono essere alti nella parte sinistra del grafico e bassi nella parte destra (se un grafico non cumulativo presenta un andamento a denti di sega, è possibile livellarlo riducendo il numero di quantili da rappresentare ed eseguendo nuovamente il grafico). La presenza di depressioni sulla sinistra del grafico o di picchi sulla destra può indicare aree in cui il modello ha una scarsa efficacia predittiva. Una linea piatta che attraversa tutto il grafico indica un modello che non fornisce informazioni utili.

Grafici dei profitti. I grafici dei profitti cumulati iniziano sempre nella parte sinistra allo 0% e terminano nella parte destra al 100%. Per un modello attendibile, il grafico dei profitti aumenterà rapidamente verso il 100% per poi scendere. Un modello che non fornisce informazioni seguirà la diagonale che va dalla parte inferiore sinistra alla parte superiore destra (tale diagonale viene visualizzata nel grafico se l'opzione **Includi linea base** è stata selezionata).

Grafici guadagno cumulativo. I grafici guadagno cumulativo tendono a iniziare nella parte sinistra al di sopra di 1,0 e quindi decrescono gradualmente verso destra fino a tornare a 1,0. Il margine destro del grafico rappresenta l'intero insieme dei dati, pertanto il rapporto dei risultati nei quantili cumulati rispetto ai risultati nei dati è 1,0. Il guadagno cumulativo di un modello efficace deve iniziare nella parte sinistra all'altezza di un valore molto più alto di 1,0, quindi rimanere a un livello elevato e infine decrescere rapidamente nella parte destra del grafico fino a 1,0. Per un modello che non fornisce informazioni, la linea si manterrà attorno a 1,0 per l'intero grafico (se è selezionata l'opzione **Includi linea base**, nel grafico viene mostrata come riferimento una linea orizzontale posizionata all'altezza di 1,0).

Grafici delle risposte. I grafici delle risposte cumulati tendono a essere molto simili ai grafici guadagno cumulativo, tranne per la graduazione. I grafici delle risposte in genere iniziano a un livello vicino al 100% e scendono gradualmente nella parte destra del grafico fino a raggiungere la percentuale di risposta globale (ovvero il risultato della divisione dei risultati totali per i record totali). La linea di un modello efficace inizierà nella parte sinistra a un livello vicino o coincidente con 100%, si manterrà a un livello elevato e quindi decrescerà rapidamente nella parte destra del grafico fino a raggiungere la percentuale di risposta globale. Per un modello che non fornisce informazioni, la linea si manterrà attorno al tasso di risposta globale per l'intero grafico (se è selezionata l'opzione **Includi linea base**, nel grafico viene mostrata come riferimento una linea orizzontale in corrispondenza del tasso di risposta globale).

Grafici dei profitti. I grafici dei profitti cumulati mostrano la somma dei profitti man mano che viene aumentata la dimensione del campione selezionato, procedendo da sinistra verso destra. I grafici dei profitti iniziano in genere vicino allo 0, aumentano costantemente man mano che ci si sposta verso destra fino a raggiungere un picco o un livello elevato costante al centro e poi scendono verso il margine destro del grafico. I profitti di un modello efficace mostreranno un picco ben definito nella parte centrale del grafico. Per un modello che non fornisce informazioni, la linea sarà relativamente diritta e potrà seguire un andamento crescente, decrescente o piatto a seconda della struttura costo/entrate pertinente.

Grafici ROI. I grafici ROI (Return On Investment) cumulati tendono a essere simili ai grafici delle risposte e ai grafici guadagno cumulativo, tranne per la graduazione. I grafici ROI in genere iniziano a un livello superiore allo 0% e decrescono gradualmente fino a raggiungere il ROI globale per l'intero insieme di dati (che può essere negativo). La linea di un modello efficace deve iniziare a un livello molto più alto dello 0%, mantenersi a un livello elevato e quindi decrescere abbastanza rapidamente nella parte destra del grafico verso il ROI globale. Per un modello che non fornisce informazioni, la linea si manterrà attorno al valore del ROI globale.

Grafici ROC. Le curve ROC in genere hanno la forma di un grafico dei profitti cumulato. La curva inizia alla coordinata (0,0) e termina alla coordinata (1,1) da sinistra a destra. Un grafico che si impenna bruscamente verso la coordinata (0,1) e poi si stabilizza indica un buon classificatore. Un modello che classifica istanze in modo casuale come risultati o mancati risultati seguirà la diagonale che va dalla parte inferiore sinistra alla parte superiore destra (visualizzata nel grafico se **Includi baseline** è selezionato). Se non viene fornito un campo di confidenza per un modello, il modello viene tracciato come singolo punto. Il classificatore con la soglia ottimale di classificazione si trova più vicino alla coordinata (0,1), oppure nell'angolo superiore sinistro del grafico. Questa posizione rappresenta un numero elevato di istanze classificate correttamente come risultati, ed un numero limitato di istanze classificate non correttamente come risultati. I punti sopra la linea della diagonale rappresentano buoni risultati di classificazione. I punti sotto la linea della diagonale rappresentano scarsi risultati di classificazione, che sono peggiori di una classificazione casuale delle istanze.

Utilizzo di un grafico di valutazione

L'utilizzo del mouse per esplorare un grafico di valutazione è simile all'utilizzo di un istogramma o di un grafico raccolto. L'asse x rappresenta i punteggi del modello dei quantili specificati, ad esempio ventili o decili.

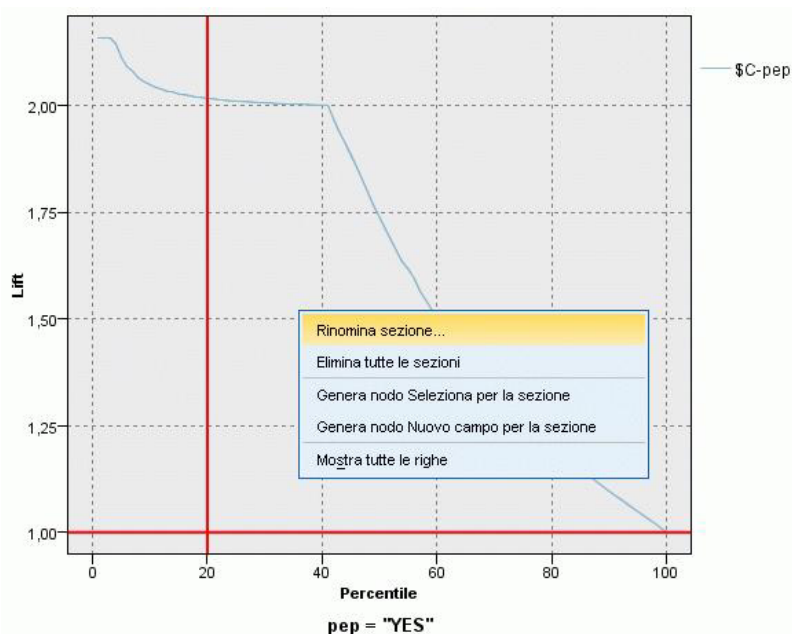


Figura 49. Utilizzo di un grafico di valutazione

È possibile suddividere l'asse x in sezioni in modo analogo a un istogramma utilizzando l'icona di divisione che fornisce opzioni per la suddivisione automatica dell'asse in sezioni uguali. Per ulteriori informazioni, consultare l'argomento "Esplorazione dei grafici". Per modificare manualmente i limiti delle sezioni, selezionare **Sezioni grafico** dal menu Modifica.

Dopo aver creato un grafico di valutazione, definito sezioni ed esaminato i risultati, è possibile utilizzare le opzioni del menu Genera e il menu di scelta rapida per creare automaticamente nodi basati sulle selezioni nel grafico. Per ulteriori informazioni, consultare l'argomento "Generazioni di nodi da grafici" a pagina 251.

Quando si generano nodi da un grafico di valutazione, viene richiesto di selezionare un singolo modello tra tutti quelli disponibili nel grafico.

Selezionare un modello e fare clic su **OK** per generare il nuovo nodo nell'area del flusso.

Esplorazione dei grafici

Mentre la modalità Modifica consente di modificare layout e aspetto del grafico, la modalità Esplora consente di esplorare analiticamente dati e valori rappresentati dal grafico. Lo scopo principale dall'esplorazione è l'analisi dei dati e la successiva identificazione di valori tramite l'utilizzo di sezioni, aree ed elementi contrassegnati per generare nodi seleziona, nuovo campo o bilanciamento. Per selezionare questa modalità, scegliere **Visualizza > Modalità Esplora** dai menu (oppure fare clic sull'icona nella barra degli strumenti).

Mentre alcuni grafici supportano tutti gli strumenti di esplorazione, altri ne accettano uno solo. La modalità Esplora consente di:

- Definire e modificare sezioni, che vengono utilizzate per suddividere i valori lungo un asse x scala. Per ulteriori informazioni, consultare l'argomento "Utilizzo delle sezioni" a pagina 244.
- Definire e modificare aree, che vengono utilizzate per identificare un gruppo di valori all'interno di un'area rettangolare. Per ulteriori informazioni, consultare l'argomento "Utilizzo delle aree" a pagina 248.

- Aggiungere ed eliminare contrassegni sugli elementi per selezionare manualmente valori che potrebbero essere utilizzati per generare un nodo Seleziona o Ricava. Per ulteriori informazioni, consultare l'argomento "Utilizzo degli elementi contrassegnati" a pagina 250.
- Generare nodi utilizzando i valori identificati da sezioni, aree, elementi contrassegnati e collegamenti Web da utilizzare nel flusso. Per ulteriori informazioni, consultare l'argomento "Generazioni di nodi da grafici" a pagina 251.

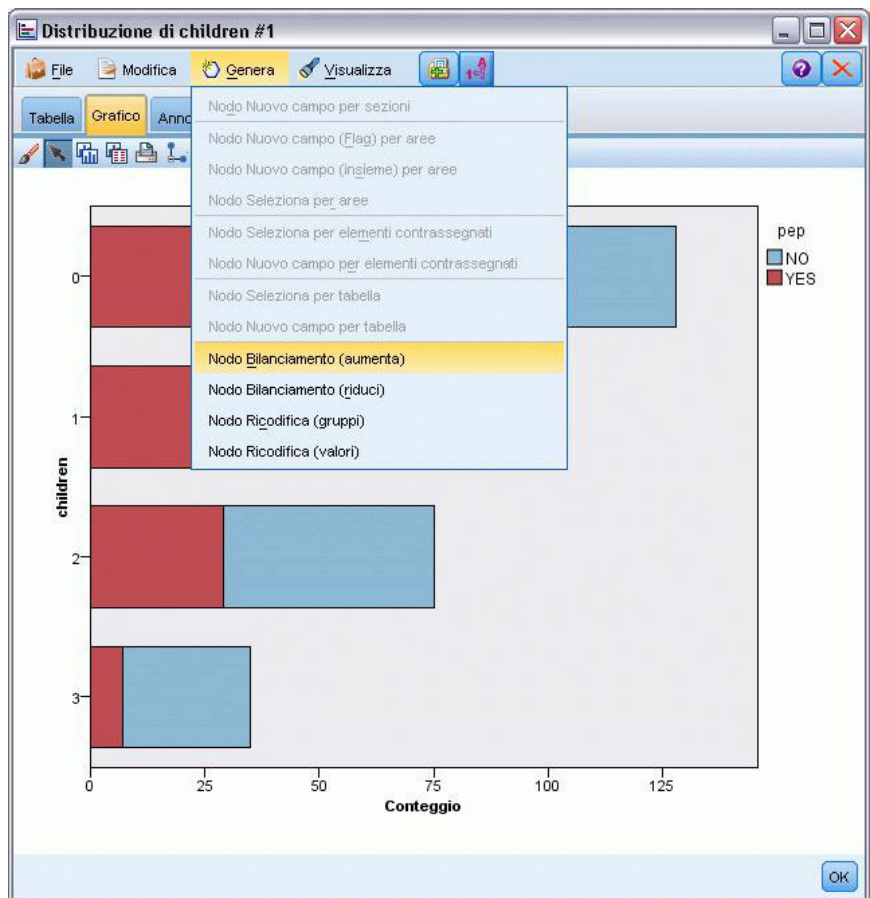


Figura 50. Grafico con il menu Genera visualizzato

Utilizzo delle sezioni

In tutti i grafici con un campo scala sull'asse x è possibile disegnare righe di sezione verticali per suddividere l'intervallo di valori sull'asse x . Se un grafico ha più pannelli, una riga di sezione tracciata in un pannello viene rappresentata anche negli altri pannelli.

Non tutti i grafici supportano le sezioni. Di seguito sono elencati alcuni di quelli che le supportano: istogrammi, grafici a barre e di distribuzione, grafici (a linee, a dispersione, tempo e così via), grafici di raccolta e di valutazione. Nei grafici con pannelli, le sezioni vengono visualizzate in tutti i pannelli. Inoltre, nei grafici SPLOM è visibile in alcuni casi una riga di sezione orizzontale, poiché l'asse su cui è stata tracciata la sezione di campo/variabile è stato capovolto.

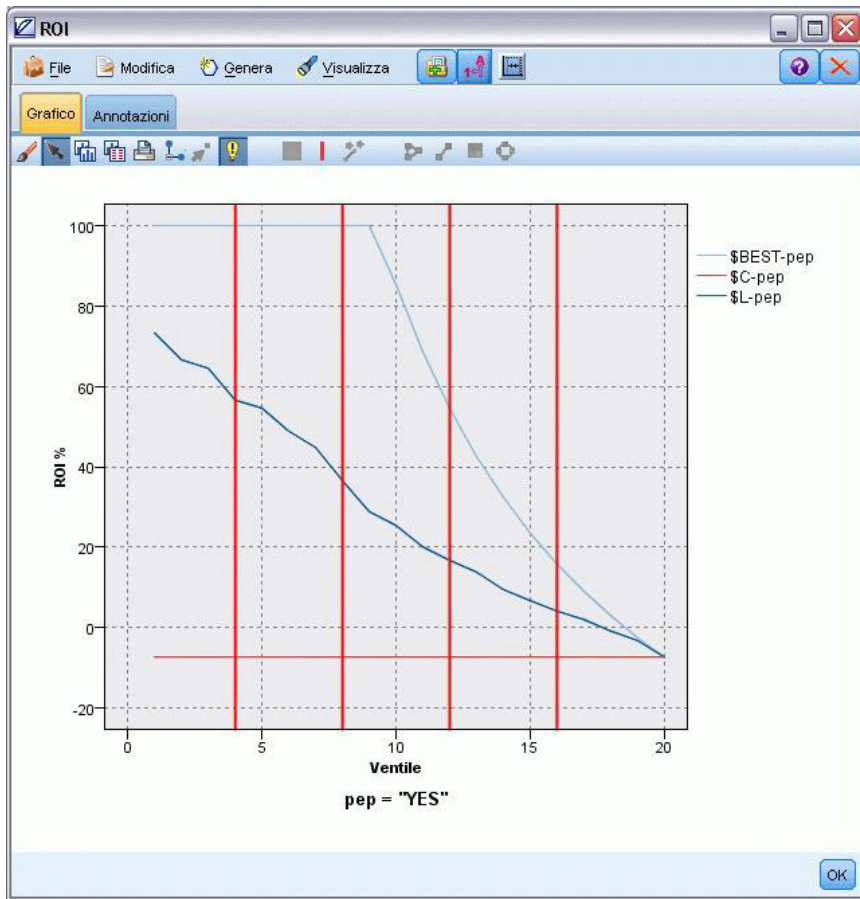


Figura 51. Grafico con tre sezioni

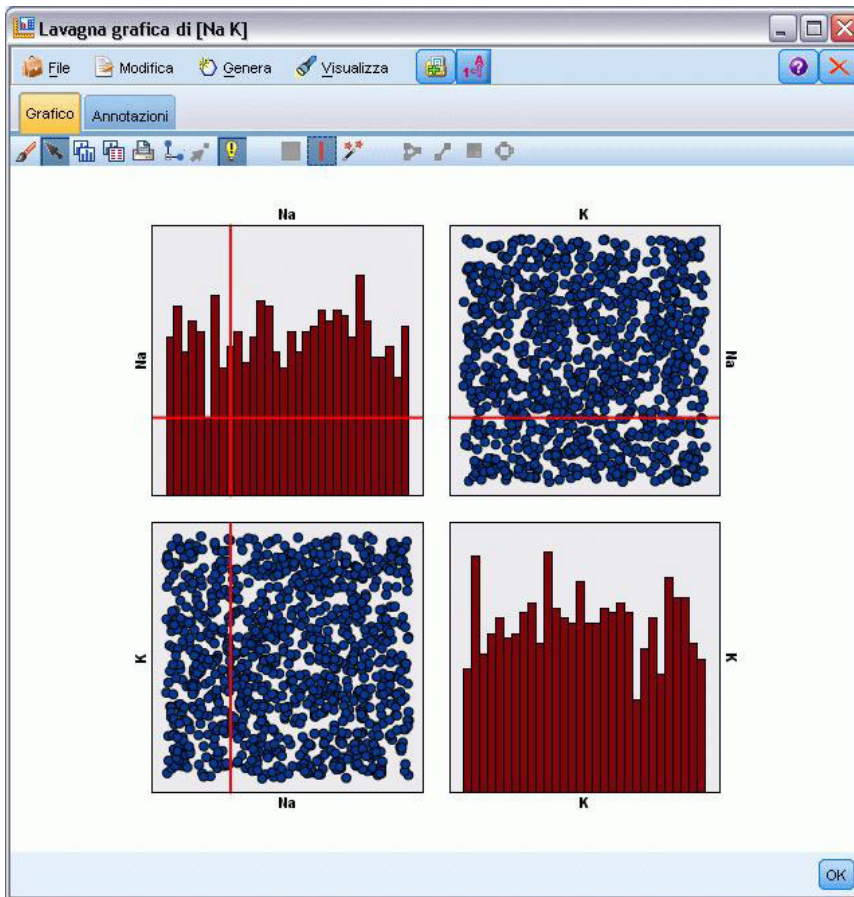


Figura 52. SPLOM con sezioni

Definizione delle sezioni

In un grafico privo di sezioni, aggiungendo una riga di sezione si suddivide il grafico in due sezioni. Il valore della riga di sezione rappresenta il punto di partenza, denominato anche limite inferiore, della seconda sezione leggendo il grafico da sinistra verso destra. Analogamente, in un grafico con due sezioni, l'aggiunta di una riga di sezione suddivide una sezione in due, generandone tre. Per default, le sezioni sono denominate *sezioneN*, dove *N* corrisponde al numero di sezioni sull'asse *x* calcolate da sinistra verso destra.

Dopo aver definito una sezione, è possibile selezionarla e trascinarla per riposizionarla sull'asse *x*. Facendo clic con il pulsante destro del mouse all'interno di una sezione sono disponibili ulteriori scelte rapide che consentono di rinominare, eliminare e generare nodi per tale sezione.

Per definire sezioni:

1. Controllare che sia attiva la modalità Esplora. Dai menu, scegliere **Visualizza > Modalità Esplora**.
2. Sulla barra degli strumenti della modalità Esplora, fare clic sul pulsante Crea sezione.



Figura 53. Pulsante della barra degli strumenti Crea sezione

3. In un grafico che supporta le sezioni, fare clic sul punto del valore dell'asse *x* in corrispondenza del quale si desidera definire una riga di sezione.

Nota: in alternativa, fare clic sull'icona della barra degli strumenti **Suddividi il grafico in sezioni**, immettere il numero di sezioni uguali desiderate e fare clic su **Suddividi**.



Figura 54. L'icona di divisione utilizzata per espandere la barra degli strumenti con opzioni relative alla suddivisione in sezioni

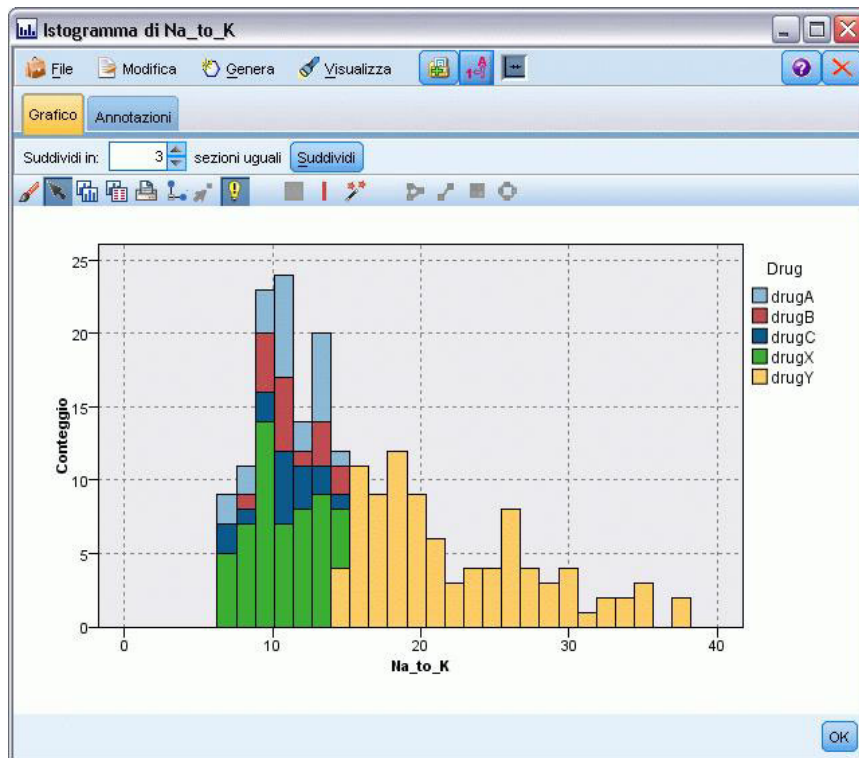


Figura 55. Creazione di sezioni uguali con le sezioni attivate

Modifica, ridenominazione ed eliminazione delle sezioni

È possibile modificare le proprietà di sezioni esistenti nella finestra di dialogo Modifica sezioni grafico oppure tramite menu di scelta rapida nel grafico stesso.

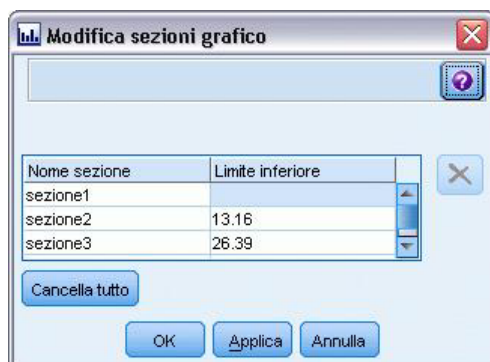


Figura 56. Finestra di dialogo Modifica sezioni grafico

Per modificare sezioni:

1. Controllare che sia attiva la modalità Esplora. Dai menu, scegliere **Visualizza > Modalità Esplora**.
2. Sulla barra degli strumenti della modalità Esplora, fare clic sul pulsante Crea sezione.
3. Dai menu, scegliere **Modifica > Sezioni grafico**. Viene visualizzata la finestra di dialogo Modifica sezioni grafico.
4. Se il grafico include più campi (per esempio i grafici SPLOM), è possibile selezionare il campo desiderato dall'elenco a discesa.
5. Per aggiungere una nuova sezione, digitarne il nome e il limite inferiore. Premere quindi INVIO per iniziare una nuova riga.
6. Per modificare il limite di una sezione, regolarne il **Limite inferiore**.
7. Per rinominare una sezione, immettere un nuovo nome di sezione.
8. Per eliminare una sezione, selezionare la riga corrispondente nella tabella e fare clic sul pulsante Elimina.
9. Fare clic su **OK** per applicare le modifiche e chiudere la finestra di dialogo.

Nota: in alternativa, è possibile eliminare e rinominare sezioni direttamente nel grafico facendo clic con il pulsante destro del mouse sulla riga della sezione e scegliendo l'opzione desiderata dai menu di scelta rapida.

Utilizzo delle aree

In tutti i grafici con due assi scala (o intervallo) è possibile disegnare aree per raggruppare valori all'interno di un'area rettangolare denominata appunto "area". Per **area** si intende una porzione del grafico delimitata da valori X e Y minimi e massimi. Se un grafico ha più pannelli, un'area tracciata in un pannello viene rappresentata anche negli altri pannelli.

Non tutti i grafici supportano le aree. Di seguito sono elencati alcuni di quelli che le supportano: grafici (a linee, a dispersione, a bolle, tempo e così via), SPLOM e grafici raccolta. Tali aree vengono disegnate nello spazio X, Y e non possono pertanto essere definite in grafici 1-D, 3-D o animati. Nei grafici con pannelli, le aree vengono visualizzate in tutti i pannelli. In un grafico SPLOM (scatterplot matrix), un'area corrispondente verrà visualizzata nei plot superiori corrispondenti ma non in quelli diagonali poiché essi mostrano solo un campo scala.

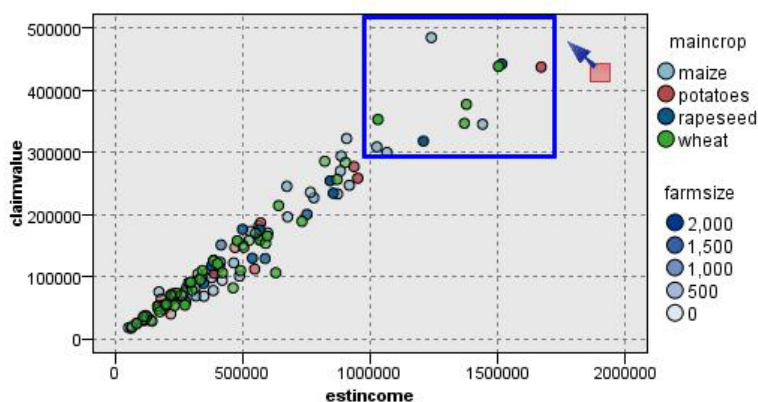


Figura 57. Definizione di un'area di valori di richiesta elevati

Definizione delle aree

Ogni volta che si definisce un'area, si crea un raggruppamento di valori. Per default, ogni nuova area viene denominata *Area<N>*, dove *N* corrisponde al numero di aree già create.

Dopo aver definito un'area, è possibile fare clic con il pulsante destro del mouse sulla riga dell'area per visualizzare alcune scelte rapide di base. Tuttavia, facendo clic con il pulsante destro del mouse all'interno dell'area (anziché sulla riga) sono disponibili ulteriori scelte rapide che consentono di rinominare, eliminare e generare nodi Seleziona e Ricava per tale area.

È possibile selezionare sottoinsiemi di record in base alla loro inclusione in una particolare area o in un'area formata da più aree. È inoltre possibile incorporare le informazioni dell'area per un record creando un nodo Ricava per generare un flag per i record in base all'inclusione in un'area. Per ulteriori informazioni, consultare l'argomento "Generazioni di nodi da grafici" a pagina 251.

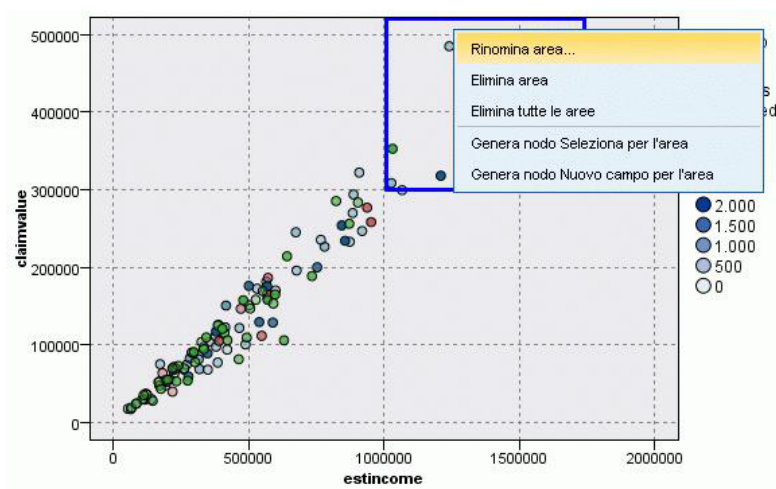


Figura 58. Analisi dell'area contenente valori di richiesta elevati

Per definire aree:

1. Controllare che sia attiva la modalità Esplora. Dai menu, scegliere **Visualizza > Modalità Esplora**.
2. Sulla barra degli strumenti della modalità Esplora, fare clic sul pulsante Crea area.



Figura 59. Pulsante della barra degli strumenti Crea area

3. In un grafico che supporta le aree, trascinare tenendo premuto il pulsante del mouse per disegnare un'area rettangolare.

Modifica, ridenominazione ed eliminazione delle aree

È possibile modificare proprietà di aree esistenti nella finestra di dialogo Modifica aree grafico oppure tramite menu di scelta rapida nel grafico stesso.



Figura 60. Impostazione delle proprietà relative alle aree definite

Per modificare aree:

1. Controllare che sia attiva la modalità Esplora. Dai menu, scegliere **Visualizza > Modalità Esplora**.
2. Sulla barra degli strumenti della modalità Esplora, fare clic sul pulsante Crea area.
3. Dai menu, scegliere **Modifica > Aree grafico**. Viene visualizzata la finestra di dialogo Modifica aree grafico.
4. Se il grafico include più campi (per esempio, grafici SPLOM), è necessario definire il campo per l'area nelle colonne *Campo A* e *Campo B*.
5. Per aggiungere una nuova area su una nuova riga, digitarne il nome, selezionare eventuali nomi di campo e definire i limiti massimo e minimo per ciascun campo. Premere quindi INVIO per iniziare una nuova riga.
6. Per modificare limiti di area esistenti, regolare i valori **Min** e **Max** per *A* e *B*.
7. Per rinominare un'area, modificarne il nome nella tabella.
8. Per eliminare un'area, selezionare la riga corrispondente nella tabella e fare clic sul pulsante Elimina.
9. Fare clic su **OK** per applicare le modifiche e chiudere la finestra di dialogo.

Nota: in alternativa, è possibile eliminare e rinominare le regioni direttamente nel grafico facendo clic con il pulsante destro del mouse sulla riga della regione e scegliendo l'opzione desiderata dai menu di scelta rapida.

Utilizzo degli elementi contrassegnati

È possibile contrassegnare elementi, quali barre, sezioni e punti di un grafico. Non è possibile contrassegnare linee, aree e superfici di grafici diversi da grafici tempo, multiplot e grafici di valutazione, poiché in questo caso le linee fanno riferimento a campi. Ogni volta che si contrassegna un elemento, si evidenziano essenzialmente tutti i dati rappresentati da tale elemento. In un grafico dove lo stesso caso è rappresentato in più posizioni (per esempio, un grafico SPLOM), contrassegnare equivale a "pennellare". È possibile contrassegnare elementi di grafici anche all'interno di sezioni e aree. Ogni volta che si contrassegna un elemento e si torna in modalità Modifica, gli elementi contrassegnati rimangono visibili.

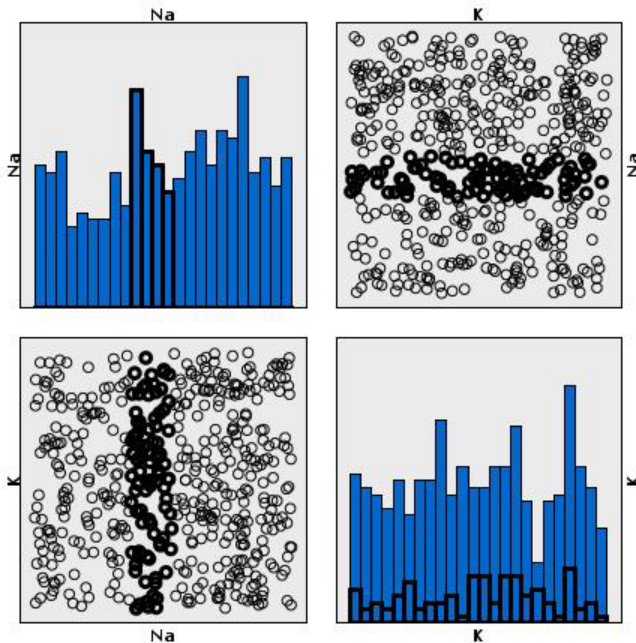


Figura 61. Contrassegno di elementi in un grafico SPLOM

È possibile contrassegnare elementi ed eliminare i contrassegni in un grafico facendo clic sugli elementi desiderati. Al primo clic, l'elemento viene contrassegnato e appare con un bordo spesso colorato a indicare che è stato contrassegnato. Facendo nuovamente clic sull'elemento, il bordo scompare e l'elemento non risulta più contrassegnato. Per contrassegnare più elementi, è possibile tenere premuto il tasto Ctrl mentre si fa clic sugli elementi desiderati oppure trascinare il mouse intorno a ciascun elemento da contrassegnare utilizzando la "bacchetta magica". Ricordare che, se si fa clic su un'altra area o un altro elemento senza tenere premuto Ctrl, tutti gli elementi precedentemente contrassegnati perdono il contrassegno.

Dagli elementi contrassegnati di un grafico è possibile generare nodi Seleziona e Ricava. Per ulteriori informazioni, consultare l'argomento "Generazioni di nodi da grafici".

Per contrassegnare elementi:

1. Controllare che sia attiva la modalità Esplora. Dai menu, scegliere **Visualizza > Modalità Esplora**.
2. Sulla barra degli strumenti della modalità Esplora, fare clic sul pulsante Contrassegna elementi.
3. Fare clic sull'elemento desiderato oppure fare clic e trascinare tenendo premuto il pulsante del mouse per delimitare l'area che contiene più elementi.

Generazioni di nodi da grafici

Una delle funzionalità più potenti offerte dai grafici di IBM SPSS Modeler è la possibilità di generare nodi da un grafico o da una selezione all'interno di un grafico. Per esempio, da un grafico tempo è possibile generare nodi Seleziona e Ricava in base a una selezione o a un'area di dati, utilizzando quindi un sottoinsieme di dati. Questa potente funzione può essere utilizzata, per esempio, per individuare ed escludere i valori anomali.

Ogni volta che si disegna una sezione, è inoltre possibile generare un nodo Ricava. Nei grafici con due assi scala è possibile generare nodi Seleziona o Ricava dalle aree disegnate nel grafico. Nei grafici con elementi contrassegnati è possibile generare nodi Ricava, nodi Selezione e in taluni casi nodi Filtro da tali elementi. È inoltre possibile generare nodi bilanciamento per tutti quei grafici che mostrano una

distribuzione di conteggi.

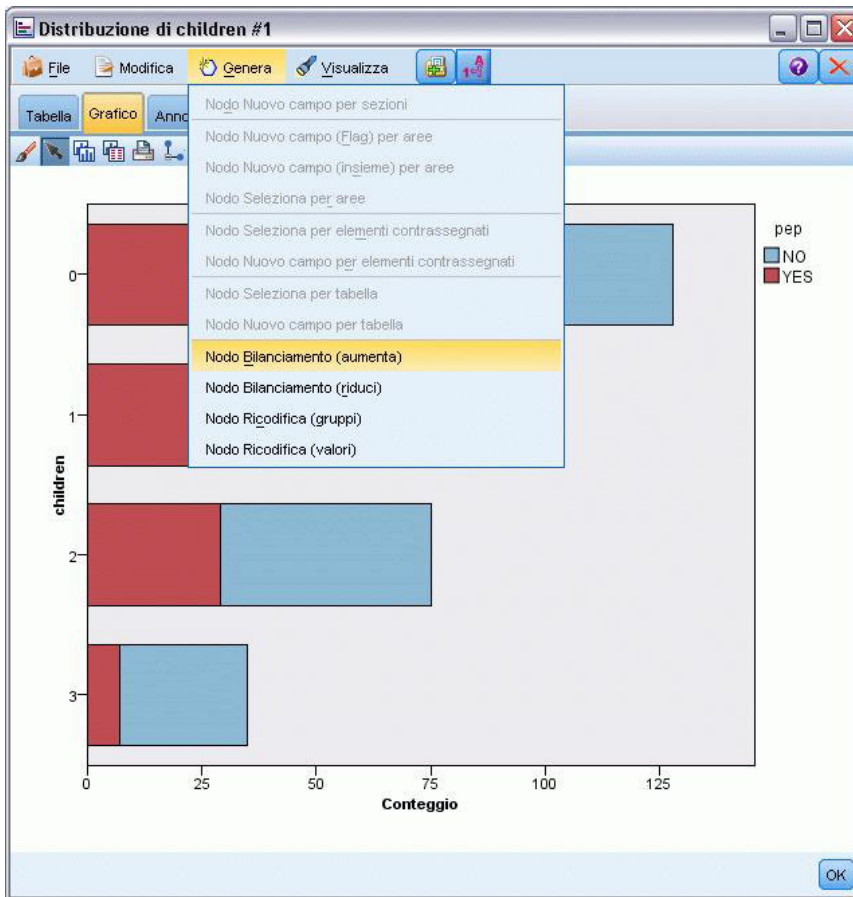


Figura 62. Grafico con il menu Genera visualizzato

Ogni volta che si genera un nodo, questo viene posizionato direttamente nell'area del flusso in modo che possa essere collegato a un flusso esistente. Dai grafici è possibile generare i seguenti nodi: seleziona, nuovo campo, bilanciamento, filtro e ricodifica.

Nodi Seleziona

È possibile generare nodi Seleziona per verificare l'inclusione dei record all'interno di una determinata area e l'esclusione di tutti gli altri record al di fuori di tale area o viceversa allo scopo di accelerare l'elaborazione.

- **Per sezioni.** È possibile generare un nodo Seleziona che includa o escluda i record all'interno di una determinata sezione. L'opzione **Nodo Seleziona solo per sezioni** è disponibile unicamente tramite il menu di scelta rapida, poiché è necessario selezionare quale sezione utilizzare nel nodo Seleziona.
- **Per aree.** È possibile generare un nodo Seleziona che includa o escluda i record all'interno di una determinata area.
- **Per elementi contrassegnati.** È possibile generare nodi Seleziona per acquisire i record corrispondenti agli elementi contrassegnati o ai collegamenti Web di un grafico.

Nodi Ricava

È possibile generare nodi Ricava da aree, sezioni ed elementi contrassegnati. Tutti i grafici possono produrre nodi Ricava. Nel caso di grafici di valutazione, viene visualizzata una finestra di dialogo per la selezione del modello. In caso di grafici Web, è possibile selezionare **Nodo Ricava ("E")** e **Nodo Ricava ("O")**.

- **Per sezioni.** È possibile generare un nodo Ricava che produce una categoria per ogni intervallo contrassegnato sull'asse, utilizzando come nomi di categoria i nomi di sezione elencati nella finestra Modifica sezioni.
- **Per aree.** È possibile generare un nodo Ricava (**Deriva come flag**) che crea un campo flag denominato *in_area* con i flag impostati su *V* per i record inclusi in un'area e su *F* per i record esterni a tutte le aree. È inoltre possibile generare un nodo Ricava (**Deriva come insieme**) che produce un insieme con un valore per ogni area con un nuovo campo denominato *area* per ogni record, che assume come proprio valore il nome dell'area contenente i record. Ai record esterni a qualsiasi area viene assegnato il nome dell'area di default I nomi dei valori diventano i nomi delle aree elencati nella finestra di dialogo Modifica aree.
- **Per elementi contrassegnati.** È possibile generare un nodo Ricava che calcola un flag che è *Vero* per tutti gli elementi contrassegnati e *Falso* per tutti gli altri record.

Nodi bilanciamento

È possibile generare nodi bilanciamento per correggere eventuali squilibri nei dati, per esempio per ridurre la frequenza di valori comuni [utilizzare la voce di menu **Nodo bilanciamento (riduci)**] o per aumentare l'occorrenza di valori non frequenti [utilizzare la voce di menu **Nodo bilanciamento (aumenta)**]. È possibile generare nodi bilanciamento per tutti quei grafici che mostrano una distribuzione di conteggi, quali istogrammi, grafici a punti, grafici raccolta, grafici con barre dei conteggi, grafici con torta dei conteggi e multiplot.

Nodi Filtro

È possibile generare nodi Filtro per rinominare o filtrare campi in base alle linee o ai nodi contrassegnati nel grafico. Nel caso dei grafici di valutazione, la linea di migliore approssimazione non genera un nodo Filtro.

Nodi Ricodifica

È possibile generare nodi Ricodifica per ricodificare valori. Questa opzione è utilizzata per i grafici di distribuzione. È possibile generare un nodo Ricodifica per **gruppi** per ricodificare valori specifici di un campo visualizzato in base alla loro inclusione in un gruppo (selezionare i gruppi con Ctrl+clic nella scheda Tabelle). È inoltre possibile generare un nodo Ricodifica per **valori** per ricodificare i dati in un insieme esistente di numerosi valori, per esempio per ricodificare i dati in un insieme standard di valori al fine di unire i dati finanziari di varie aziende per l'analisi.

Nota: se i valori sono predefiniti, è possibile leggerli in IBM SPSS Modeler come un file flat e utilizzare un nodo distribuzione per visualizzare tutti i valori. Generare quindi un nodo Ricodifica (valori) per questo campo direttamente dal grafico. In questo modo tutti i valori obiettivo verranno inseriti nella colonna *Nuovi valori* del nodo Ricodifica (elenco a discesa).

Generazioni di nodi da grafici

È possibile utilizzare il menu Genera della finestra di output del grafico per generare nodi. Il nodo generato verrà posizionato nell'area del flusso. Per utilizzare il nodo è necessario connetterlo a un flusso esistente.

Per generare un nodo da un grafico:

1. Controllare che sia attiva la modalità Esplora. Dai menu, scegliere **Visualizza > Modalità Esplora**.
2. Sulla barra degli strumenti della modalità Esplora, fare clic sul pulsante Area.

3. Definire eventuali sezioni, aree o elementi contrassegnati necessari per generare il nodo.
4. Dal menu Genera, scegliere il tipo di nodo che si desidera generare. Sono attivati solo quelli che è possibile generare.

Nota: in alternativa, è inoltre possibile generare nodi direttamente dal grafico facendo clic con il pulsante destro del mouse e scegliendo l'opzione Genera desiderata dai menu di scelta rapida.

Modifica delle visualizzazioni

Mentre la modalità Esplora consente di esaminare in modo analitico i dati e i valori presenti nella visualizzazione, con la modalità Modifica è possibile intervenire sul layout e sull'aspetto della visualizzazione. Per esempio, è possibile modificare caratteri e colori in base alla guida di stile della propria organizzazione. Per selezionare questa modalità, scegliere **Visualizza > Modalità Modifica** dai menu (oppure facendo clic sull'icona nella barra degli strumenti).

In modalità Modifica sono disponibili numerose barre degli strumenti che riguardano diversi aspetti del layout della visualizzazione. Se lo si desidera, è possibile nascondere le barre degli strumenti che non vengono utilizzate in modo da incrementare lo spazio della finestra di dialogo in cui il grafico viene visualizzato. Per selezionare o annullare la selezione delle barre degli strumenti, fare clic sui relativi nomi nel menu Visualizza.

Nota: per aggiungere ulteriori dettagli alle proprie visualizzazioni, è possibile applicare titolo, note a piè di pagina ed etichette degli assi. Per ulteriori informazioni, consultare l'argomento "Aggiunta di titoli e annotazioni" a pagina 266.

In **modalità Modifica** sono disponibili svariate opzioni per modificare le visualizzazioni. È possibile:

- Modificare il testo e formattarlo.
- Modificare il colore di riempimento, la trasparenza e il motivo delle cornici e degli elementi grafici.
- Modificare il colore e il tratteggio dei bordi e delle linee.
- Ruotare e modificare la forma e le proporzioni degli elementi puntiformi.
- Modificare le dimensioni degli elementi grafici (quali barre e punti).
- Regolare lo spazio attorno agli elementi utilizzando margini e padding.
- Specificare la formattazione dei numeri.
- Modificare le impostazioni degli assi e delle scale.
- Ordinare, escludere e comprimere categorie su un asse di categorie.
- Impostare l'orientamento dei riquadri.
- Applicare trasformazioni a un sistema di coordinate.
- Modificare le statistiche, i tipi di elemento grafico e i modificatori di collisione.
- Modificare la posizione della legenda.
- Applicare fogli di stile di visualizzazione.

Le sezioni seguenti illustrano come eseguire le attività elencate sopra. Si consiglia inoltre di leggere le regole generali per la modifica dei grafici.

Come passare alla modalità Modifica

Dai menu, scegliere:

Visualizza > Modalità Modifica

Regole generali per la modifica delle visualizzazioni

Modalità Modifica

Tutte le modifiche vengono apportate in modalità Modifica, attivabile nei menu selezionando:

Visualizza > Modalità Modifica

Selezione

Le opzioni di modifica disponibili dipendono dalla selezione. A seconda dell'elemento selezionato, nella barra degli strumenti e nella palette delle proprietà vengono attivate opzioni diverse. Per la selezione corrente sono valide solo le opzioni attivate. Per esempio, se si seleziona un asse, nella palette delle proprietà saranno disponibili le schede Scala, Segni di riferimento maggiori e Segni di graduazione minori.

Di seguito vengono riportati suggerimenti per la selezione degli elementi nella visualizzazione:

- Fare clic su un elemento per selezionarlo.
- Selezionare un elemento grafico (quale i punti di un grafico a dispersione o le barre di un grafico a barre) con un singolo clic. Dopo questa prima selezione, fare nuovamente clic per restringere la selezione a gruppi di elementi grafici o a un singolo elemento grafico.
- Premere Esc per deselezionare tutto.

Palette

Quando nella visualizzazione è selezionato un oggetto, le diverse palette vengono aggiornate per riflettere tale selezione. Le palette contengono i controlli per apportare modifiche alla selezione. Possono essere barre degli strumenti o un riquadro con più comandi e più schede. Poiché le palette possono essere nascoste, è opportuno verificare che siano visibili quelle necessarie per apportare le modifiche. Tale verifica è possibile utilizzando il menu Visualizza.

Le palette possono essere spostate facendo clic e trascinando lo spazio vuoto all'interno della palette di una barra degli strumenti o il bordo sinistro di altre palette. Indicazioni visive mostrano il punto in cui è possibile agganciare la palette. Nel caso di palette non relative a una barra degli strumenti, è possibile fare clic sul pulsante di chiusura per nascondere la palette e quindi su quello di sgancio per visualizzarla in un'altra finestra. Il pulsante Aiuto consente di visualizzare informazioni sulla palette in questione.

Impostazioni automatiche

Alcune impostazioni sono dotate di un'opzione **-automatico-** che indica l'applicazione di valori automatici. Le impostazioni automatiche utilizzate dipendono dalla specifica visualizzazione e dai valori dei dati. È possibile inserire un valore e ignorare l'impostazione automatica. Per ripristinare l'impostazione automatica, eliminare il valore corrente e premere Invio. Per l'impostazione sarà visualizzata nuovamente la dicitura **-automatico-**.

Rimuovere o nascondere elementi

È possibile rimuovere/nascondere vari elementi nella visualizzazione. Per esempio, è possibile nascondere la legenda o l'etichetta dell'asse. Per eliminare un elemento, selezionarlo e premere Canc. Se l'eliminazione dell'elemento non è consentita, l'operazione non avrà alcun effetto. Se si elimina accidentalmente un elemento, premere Ctrl+Z per annullare l'eliminazione.

State

Alcune barre degli strumenti cambiano a seconda dello stato della selezione corrente, mentre altre no. La palette delle proprietà rispecchia sempre lo stato. Se una barra degli strumenti *non* rispecchia lo stato della selezione, questo è segnalato nell'argomento che descrive la barra degli strumenti.

Modifica e formattazione del testo

È possibile modificare il testo direttamente nel grafico e cambiare la formattazione di un intero blocco di testo. Si noti che non è possibile modificare il testo collegato direttamente ai valori dei dati. Per esempio, non è possibile modificare l'etichetta di tratto poiché il suo contenuto deriva dai dati sottostanti. È comunque possibile formattare tutto il testo della visualizzazione.

Come modificare il testo direttamente nel grafico

1. Fare doppio clic sul blocco di testo, per selezionare tutto il testo. Tutte le barre degli strumenti sono disattivate, dal momento che non si può intervenire su altre parti della visualizzazione durante la modifica del testo.
2. Digitare il testo con cui sostituire quello selezionato. È possibile anche fare nuovamente clic sul testo per visualizzare un cursore, Posizionare il cursore nel punto desiderato ed inserire il testo aggiuntivo.

Come formattare il testo

1. Selezionare la cornice contenente il testo, senza fare doppio clic sul testo.
2. Formattare il testo mediante la barra degli strumenti dei caratteri. Se la barra degli strumenti non è attivata, verificare che sia selezionata solo la *cornice* contenente il testo. Se è selezionato il testo, la barra degli strumenti sarà disattivata.

Per i caratteri è possibile modificare:

- Colore
- Famiglia (per esempio Arial o Verdana)
- Dimensioni (l'unità è pt a meno che non ne venga indicata una diversa, per esempio pc)
- Peso
- Allineamento rispetto alla cornice del testo

La formattazione viene applicata a tutto il testo all'interno della cornice. Non è possibile modificare la formattazione delle singole lettere o parole all'interno di un determinato blocco di testo.

Modifica dei colori, dei motivi, dei tratteggi e della trasparenza

Molti degli elementi di una visualizzazione sono caratterizzati da un bordo e da un riempimento. L'esempio più ovvio sono le barre all'interno dei grafici a barre. Il colore delle barre è il colore di riempimento. Attorno alle barre può anche essere presente un bordo nero continuo.

Nella visualizzazione sono presenti elementi meno ovvi dotati di colori di riempimento. Se il colore di riempimento è trasparente è possibile che non ci si accorga della sua presenza. Si pensi per esempio al testo di un'etichetta dell'asse: il testo sembra "mobile", ma in realtà è contenuto in una cornice con colore di riempimento trasparente. La cornice diventa visibile se si seleziona l'etichetta dell'asse.

Tutte le cornici della visualizzazione possono avere uno stile di riempimento e di bordo, compresa quella che racchiude l'intera rappresentazione. Inoltre, a tutti i riempimenti è associato un livello di opacità/trasparenza che può essere regolato.

Come modificare i colori, i motivi, il tratteggio e la trasparenza

1. Selezionare l'elemento che si desidera formattare. Per esempio, selezionare le barre di un grafico a barre o una cornice contenente del testo. Se la visualizzazione è suddivisa in base a una variabile o a un campo categoriale, è anche possibile selezionare il gruppo che corrisponde a una singola categoria.

Ciò consente di modificare l'aspetto di default assegnato a quel gruppo. Per esempio, è possibile modificare il colore di uno dei gruppi sovrapposti in un grafico a barre sovrapposte.

2. Per modificare il colore di riempimento o del bordo o il motivo di riempimento, utilizzare la barra degli strumenti dei colori.

Nota: questa barra degli strumenti non rispecchia lo stato della selezione corrente.

Per modificare un colore o un riempimento, è possibile fare clic sul pulsante per selezionare l'opzione visualizzata o sulla freccia a discesa per sceglierne un'altra. Per i colori si noterà la presenza di un colore che sembra bianco attraversato da una riga diagonale rossa. Questo è il colore trasparente, che si può utilizzare, per esempio, per nascondere i bordi delle barre di un istogramma.

- Il primo pulsante controlla il colore di riempimento. Se il colore è associato ad un campo continuo o ordinale, questo pulsante modifica il colore di riempimento per il colore associato al valore più alto nei dati. È possibile utilizzare la scheda Colore nella tavolozza delle proprietà per modificare il colore associato al valore più basso ed ai dati mancanti. Il colore degli elementi viene modificato in modo incrementale dal colore Basso al colore Alto con l'aumentare dei valori dei dati sottostanti.
 - Il secondo controlla il colore del bordo.
 - Il terzo controlla il motivo di riempimento. Il motivo di riempimento utilizza il colore del bordo e quindi è visibile solo se è presente un colore del bordo visibile.
 - Il quarto controllo è un dispositivo di scorrimento e una casella di testo che controllano l'opacità, nonché il colore e il motivo di riempimento. Minore è la percentuale, minore risulta l'opacità e maggiore la trasparenza. Una percentuale pari a 100% conferisce un'opacità totale (senza trasparenza).
3. Per modificare il tratteggio di un bordo o di una linea, utilizzare la barra degli strumenti delle linee.
- Nota:* questa barra degli strumenti non rispecchia lo stato della selezione corrente.
- Come avviene per l'altra barra degli strumenti, è possibile fare clic sul pulsante per selezionare l'opzione visualizzata oppure fare clic sulla freccia a discesa per scegliere un'altra opzione.

Rotazione e modifica della forma e delle proporzioni degli elementi puntiformi

È possibile ruotare gli elementi puntiformi, assegnare loro una forma predefinita diversa o modificarne le proporzioni (il rapporto tra larghezza e altezza).

Come modificare gli elementi puntiformi

1. Selezionare gli elementi puntiformi. Non è possibile ruotare o modificare la forma e le proporzioni di un singolo elemento puntiforme.
2. Per modificare i punti, utilizzare la barra degli strumenti dei simboli.
 - Il primo pulsante consente di modificare la forma dei punti. Fare clic sulla freccia a discesa e selezionare una forma predefinita.
 - Il secondo pulsante consente di ruotare i punti secondo un orientamento specifico una bussola simbolica. Fare clic sulla freccia a discesa e trascinare l'ago nella posizione desiderata.
 - Il terzo pulsante consente di modificare le proporzioni. Fare clic prima sulla freccia a discesa e poi sul rettangolo così visualizzato, quindi trascinarlo per modificarne le proporzioni. La forma del rettangolo rappresenta le proporzioni dei punti.

Modifica delle dimensioni degli elementi grafici

È possibile modificare le dimensioni degli elementi grafici della visualizzazione, (barre, linee, punti e simili). Se le dimensioni di un elemento grafico sono determinate da una variabile o da un campo, le dimensioni specificate sono quelle *minime*.

Come modificare le dimensioni degli elementi grafici

1. Selezionare gli elementi grafici da ridimensionare.

- Utilizzare il dispositivo di scorrimento oppure inserire dimensioni specifiche per l'opzione disponibile nella barra degli strumenti dei simboli. L'unità è il pixel a meno che non ne venga indicata una diversa (vedere di seguito per l'elenco completo delle abbreviazioni di unità). È possibile anche specificare una percentuale (per esempio il 30%) per indicare che un elemento grafico utilizza la percentuale impostata dello spazio disponibile. Lo spazio disponibile dipende dal tipo di elemento grafico e dalla specifica visualizzazione.

Tabella 46. Abbreviazioni di unità valide

Abbreviazione	Unità
cm	centimetro
in	pollice
mm	millimetro
pc	pica
pt	punto
px	pixel

Impostazione dei margini e del padding

Se c'è troppo (o troppo poco) spazio intorno o all'interno di una cornice di una visualizzazione, è possibile modificarne le impostazioni dei margini e di padding. Il **margin** è la quantità di spazio compresa tra la cornice e gli elementi che la circondano. Il **padding** è la quantità di spazio compresa tra il bordo della cornice e il suo *contenuto*.

Come impostare i margini e il padding

- Selezionare la cornice di cui si desidera impostare i margini e il padding. Può trattarsi di una cornice di testo, della cornice attorno alla legenda o anche della cornice dei dati che visualizza gli elementi grafici (quali barre e punti).
- Per specificare le impostazioni, utilizzare la scheda Margini nella palette delle proprietà. Le dimensioni sono espresse in pixel a meno che non venga indicata un'unità diversa, per esempio cm o in.

Formattazione dei numeri

È possibile specificare il formato dei numeri nelle etichette di tratto su un asse continuo o nelle etichette dei valori dei dati caratterizzate da un numero. Ad esempio, è possibile specificare che i numeri visualizzati nelle etichette di tratto siano rappresentati in migliaia.

Come specificare i formati dei numeri

- Selezionare le etichette di tratto dell'asse continuo o le etichette dei valori dei dati, qualora contengano numeri.
- Fare clic sulla scheda **Formato** della palette delle proprietà.
- Selezionare le opzioni di formattazione dei numeri desiderate:

Prefisso. Un carattere da visualizzare all'inizio del numero. Immettere ad esempio il simbolo di euro (€) nel caso in cui i numeri si riferiscano ai salari in questa valuta.

Suffisso. Un carattere da visualizzare alla fine del numero. Ad esempio, immettere il simbolo di percentuale (%) nel caso in cui i numeri si riferiscano a percentuali.

Min. Cifre intere. Il numero minimo di cifre da visualizzare nella parte intera di una rappresentazione decimale. Se il valore effettivo non contiene il numero minimo di cifre, la parte intera del valore verrà riempita di zeri.

Max. Cifre intere. Il numero massimo di cifre da visualizzare nella parte intera di una rappresentazione decimale. Se il valore effettivo supera il numero massimo di cifre, la parte intera del valore verrà sostituita da asterischi.

Min. Cifre decimali. Il numero minimo di cifre da visualizzare nella parte decimale di una rappresentazione decimale o scientifica. Se il valore effettivo non contiene il numero minimo di cifre, la parte decimale del valore verrà riempita di zeri.

Max. Cifre decimali. Il numero massimo di cifre da visualizzare nella parte decimale di una rappresentazione decimale o scientifica. Se il valore effettivo supera il numero massimo di cifre, il numero decimale viene arrotondato al numero di cifre appropriato.

Scientifica. Consente di scegliere se visualizzare i numeri in notazione scientifica. La notazione scientifica è utile per numeri molto grandi o molto piccoli. **-auto-** consente all'applicazione di determinare quando è opportuno utilizzare la notazione scientifica.

Graduazione. Un fattore di scala, ossia un numero per il quale viene diviso il valore originario. Utilizzare un fattore di scala quando i numeri sono grandi ma non si desidera che l'etichetta venga eccessivamente ampliata per adattarsi al numero. Se si modifica il formato del numero delle etichette di tratto, accertarsi di modificare il titolo dell'asse per indicare come deve essere interpretato il numero. Si supponga, ad esempio, che sull'asse dei valori siano visualizzati gli stipendi e che le etichette riportino 30.000, 50.000 e 70.000. È possibile utilizzare un fattore di scala pari a 1000 per visualizzare 30, 50 e 70. Sarà quindi necessario modificare il titolo dell'asse dei valori aggiungendo il testo in migliaia.

Parentesi per -vo. Consente di scegliere se racchiudere i valori negativi tra parentesi.

Raggruppamento. Consente di scegliere se visualizzare un carattere tra gruppi di cifre. Il carattere utilizzato per il raggruppamento delle cifre è determinato dalle impostazioni internazionali in uso nel computer.

Modifica delle impostazioni degli assi e delle scale

Esistono varie opzioni per la modifica degli assi e delle scale.

Modalità di modifica delle impostazioni di scala ed assi

1. Selezionare un qualsiasi elemento dell'asse (per esempio l'etichetta dell'asse o le etichette di tratto).
2. Modificare le impostazioni dell'asse e della scala mediante le schede Scala, Segni di riferimento maggiori e Segni di graduazione minori nella palette delle proprietà.

Scheda Scala

Nota: la scheda Scala non viene visualizzata per i grafici in cui i dati sono pre-aggregati (ad esempio, gli istogrammi).

Tipo. Indica se la scala è lineare o trasformata. Le trasformazioni di scala facilitano la comprensione dei dati o la formulazione dei presupposti necessari per l'inferenza statistica. Nei grafici a dispersione si può utilizzare una scala trasformata se la relazione tra variabili o campi indipendenti e dipendenti è non lineare. Le trasformazioni di scala si possono utilizzare anche per migliorare la simmetria di un istogramma asimmetrico al fine di renderlo simile a una distribuzione normale. Si noti che la trasformazione riguarda solo la scala con cui vengono rappresentati i dati, non i dati effettivi.

- **lineare.** Indica una scala lineare, non trasformata.
- **logaritmo.** Specifica una scala trasformata con un logaritmo in base 10. Per considerare anche i valori zero e negativi, questa trasformazione utilizza una versione modificata della funzione logaritmica. Questa funzione "logaritmica sicura" viene definita come $\text{sign}(x) * \log(1 + \text{abs}(x))$. Quindi, $\text{safeLog}(-99)$ è uguale a:

$$\text{sign}(-99) * \log(1 + \text{abs}(-99)) = -1 * \log(1 + 99) = -1 * 2 = -2$$

- **potenza.** Specifica una scala trasformata mediante potenza utilizzando un esponente di 0.5. Per considerare anche i valori negativi, questa trasformazione utilizza una versione modificata della funzione potenza. Questa funzione "potenza sicura" viene definita come $\text{sign}(x) * \text{pow}(\text{abs}(x), 0.5)$. Quindi, $\text{safePower}(-100)$ è uguale a:

$$\text{sign}(-100) * \text{pow}(\text{abs}(-100), 0.5) = -1 * \text{pow}(100, 0.5) = -1 * 10 = -10$$

Min/Max/Inferiore preciso/Superiore preciso. Specificano l'intervallo della scala. Se si seleziona **Inferiore preciso** e **Superiore preciso**, l'applicazione è in grado di selezionare una scala adeguata in base ai dati. Il minimo e il massimo sono "precisi" perché in genere sono valori interi maggiori o minori dei valori dei dati massimo e minimo. Per esempio, se l'intervallo di dati è compreso fra 4 e 92, i valori inferiore e superiore preciso per la scala possono essere 0 e 100 anziché il minimo e il massimo dei dati effettivi. Prestare attenzione a non impostare un intervallo troppo piccolo che nasconda elementi importanti. Notare inoltre che non è possibile impostare un minimo e un massimo espliciti se è selezionata l'opzione **Includi zero**.

Margine inferiore/Margine superiore. Consente di creare margini all'estremità inferiore e/o superiore dell'asse. Il margine risulta perpendicolare all'asse selezionato. L'unità è il pixel a meno che non ne venga indicata una diversa, per esempio cm o in. Per esempio, se si imposta il **Margine superiore** su 5 per l'asse verticale, lungo la parte superiore della cornice dei dati correrà un margine orizzontale di 5 pixel.

Capovolgi. Specifica se la scala viene capovolta.

Includi zero. Indica che la scala deve includere lo 0. Questa opzione si utilizza in genere per i grafici a barre per fare in modo che le barre inizino in corrispondenza dello 0 anziché di un valore prossimo all'altezza della barra più piccola. Se è selezionata questa opzione, le opzioni **Min** e **Max** sono disattivate poiché non è possibile impostare un minimo e un massimo personalizzati per l'intervallo della scala.

Schede Segni di riferimento maggiori/Segni di riferimento minori

I segni di graduazione o **segni del tratto** sono le linee riportate su un asse, che indicano i valori in corrispondenza di intervalli o categorie specifiche. I **segni di riferimento maggiori** sono i segni del tratto dotati di etichette e sono anche più lunghi rispetto agli altri segni del tratto. I **segni di graduazione minori** sono i segni del tratto che compaiono tra un segno di riferimento maggiore e l'altro. Alcune opzioni sono specifiche per un determinato tipo di segno di graduazione, ma la maggior parte è disponibile sia per i segni di riferimento maggiori che per quelli minori.

Mostra segni di graduazione. Specifica se in un grafico vengono visualizzati i segni di riferimento maggiori o minori.

Mostra linee della griglia. Specifica se in corrispondenza dei segni di graduazione maggiori o minori vengono visualizzate le linee della griglia. Le **linee della griglia** sono linee che attraversano tutto il grafico da un asse all'altro.

Posizione. Specifica la posizione dei segni del tratto rispetto all'asse.

Lunghezza. Specifica la lunghezza dei segni del tratto. L'unità è il pixel a meno che non ne venga indicata una diversa, per esempio cm o in.

Base. *Viene applicata solo ai segni di riferimento maggiori.* Specifica il valore in corrispondenza del quale è visualizzato il primo segno di riferimento maggiore.

Delta. *Viene applicata solo ai segni di riferimento maggiori.* Specifica la differenza tra i segni di riferimento maggiori. Questo significa che tali segni saranno visualizzati ogni n valori, dove n è il valore delta.

Divisioni. *Viene applicata solo ai segni di riferimento minori.* Specifica il numero di divisioni con segni di riferimento minori tra i segni di riferimento maggiori. Il numero dei segni di graduazione minori è

inferiore di un'unità rispetto al numero delle divisioni. Ad esempio, si supponga che siano presenti segni di riferimento maggiori in 0 e 100. Se si immette 2 come numero di divisioni del segno di riferimento minore, sarà presente *un* segno di riferimento minore in 50, che divide l'intervallo 0-100 e crea *due* divisioni.

Modifica delle categorie

È possibile modificare le categorie su un asse di categorie in diversi modi:

- Modificare il criterio di ordinamento di visualizzazione delle categorie.
- Escludere categorie specifiche.
- Aggiunta di una categoria che non viene visualizzata nell'insieme di dati.
- Compressione/combinazione di categorie di dimensioni ridotte in un'unica categoria.

Come modificare il criterio di ordinamento delle categorie

1. Selezionare un'asse delle categorie. La palette Categorie visualizza le categorie sull'asse.

Nota: se la tavolozza non è visibile, verificare che sia abilitata. Dal menu Visualizza in IBM SPSS Modeler, selezionare **Categorie**.

2. Nella palette Categorie, selezionare un'opzione di ordinamento dall'elenco a discesa:

Personalizzato. Ordina le categorie in base all'ordine con cui appaiono nella palette. Utilizzare i pulsanti freccia per spostare le categorie in cima, su, giù e al fondo dell'elenco.

Dati. Ordina le categorie in base all'ordine con cui si presentano nell'insieme di dati.

Nome. Ordina le categorie in ordine alfabetico, utilizzando i nomi visualizzati nella palette. Questi potrebbero essere un valore oppure un'etichetta, a seconda che sia selezionato il pulsante della barra degli strumenti per visualizzare valori o etichette.

Valore. Ordina le categorie in base al valore dei dati sottostanti, utilizzando i valori visualizzati tra parentesi nella tavolozza. Questa opzione è supportata unicamente dalle sorgenti dati con metadati (ad esempio i file di dati IBM SPSS Statistics).

Statistica. Ordina le categorie in base alla statistica calcolata per ogni categoria. Esempi di statistica includono conteggi, percentuali e medie. Questa opzione è disponibile solo se nel grafico è utilizzata una statistica.

Come aggiungere una categoria

Per impostazione predefinita, sono disponibili solo le categorie visualizzate nell'insieme di dati. È possibile aggiungere una categoria alla visualizzazione, se necessario.

1. Selezionare un'asse delle categorie. La palette Categorie visualizza le categorie sull'asse.

Nota: se la tavolozza non è visibile, verificare che sia abilitata. Dal menu Visualizza in IBM SPSS Modeler, selezionare **Categorie**.

2. Nella palette Categorie, fare clic sul pulsante Aggiungi categoria:



Figura 63. Pulsante Aggiungi categoria

3. Nella finestra di dialogo Aggiungi una nuova categoria, specificare un nome per la categoria.
4. Fare clic su **OK**.

Come escludere categorie specifiche

1. Selezionare un'asse delle categorie. La palette Categorie visualizza le categorie sull'asse.

Nota: se la tavolozza non è visibile, verificare che sia abilitata. Dal menu Visualizza in IBM SPSS Modeler, selezionare **Categorie**.

2. Nella palette **Categorie**, selezionare il nome di una categoria nell'elenco **Includi**, quindi fare clic sul pulsante **X**. Per ripristinare la categoria, selezionarne il nome nell'elenco **Escludi**, quindi fare clic sulla freccia a destra dell'elenco.

Come comprimere/combinare categorie di dimensioni ridotte

È possibile raggruppare categorie di dimensioni così ridotte da non richiedere una visualizzazione separata. Se, ad esempio, si utilizza un grafico a torta composto da molte categorie, è possibile comprimerle con una percentuale inferiore a 10. La compressione delle categorie è disponibile solo per le statistiche additive. Ad esempio, non è possibile aggiungere insieme le medie poiché le medie non sono additive. Pertanto, non è disponibile la compressione/combinazione delle categorie utilizzando una media.

1. Selezionare un'asse delle categorie. La palette **Categorie** visualizza le categorie sull'asse.
Nota: se la tavolozza non è visibile, verificare che sia abilitata. Dal menu **Visualizza** in **IBM SPSS Modeler**, selezionare **Categorie**.
2. Nella palette **Categorie**, selezionare **Comprimi** e specificare una percentuale. Le categorie la cui percentuale del totale è inferiore al numero specificato vengono combinate in un'unica categoria. La percentuale è basata sulla statistica mostrata nel grafico. La compressione è disponibile solo per statistiche basate su conteggio o somma.

Modifica dei riquadri di orientamento

Se nella visualizzazione si utilizzano i riquadri, è possibile modificarne l'orientamento.

Come modificare l'orientamento dei riquadri

1. Selezionare una parte qualsiasi della visualizzazione.
2. Fare clic sulla scheda **Riquadri** nella palette delle proprietà.
3. Selezionare un'opzione nell'elenco **Layout**:

Tabella. Dispone i riquadri come una tabella, con una riga o una colonna assegnata a ogni singolo valore.

Trasposto. Dispone i riquadri come una tabella ma in più scambia fra loro le righe e le colonne originali. Questa opzione non equivale a trasporre il grafico in sé. Si noti che l'asse delle x e quello delle y restano invariati quando si seleziona questa opzione.

Elenco. Dispone i riquadri come un elenco in cui ogni cella rappresenta una combinazione di valori. Le colonne e le righe non sono più assegnate a valori singoli. Questa opzione consente ai riquadri di andare a capo, se necessario.

Trasformazione del sistema di coordinate

Molte visualizzazioni vengono rappresentate in un sistema di coordinate piatto di forma rettangolare. Tale sistema può essere trasformato in base alle proprie esigenze. È ad esempio possibile applicarvi una trasformazione polare, aggiungere effetti di ombreggiatura obliqui e trasporre gli assi. Se già applicate alla visualizzazione corrente, tutte queste trasformazioni possono anche essere annullate. Si supponga, ad esempio, di generare un grafico a torta in un sistema di coordinate polari. È possibile annullare la trasformazione polare e visualizzare il grafico a torta come singola barra sovrapposta in un sistema di coordinate rettangolare.

Come trasformare il sistema di coordinate

1. Selezionare il sistema di coordinate da trasformare facendo clic sulla cornice intorno al singolo grafico.
2. Fare clic sulla scheda **Coordinate** nella palette delle proprietà.
3. Selezionare le trasformazioni da applicare al sistema di coordinate. Una trasformazione può anche essere deselezionata per annullarla.

Trasposto. La modifica dell'orientamento degli assi, nota come **trasposizione**, Si tratta di un'operazione analoga all'inversione dell'asse verticale e di quello orizzontale in una visualizzazione 2-D.

Polare. La trasformazione polare consente di disegnare gli elementi grafici a una determinata angolazione e distanza rispetto al centro del grafico. Un grafico a torta è una visualizzazione ad una dimensione con una trasformazione polare che traccia le singole barre a specifiche angolazioni. Un grafico a radar è una visualizzazione 2D con una trasformazione polare che traccia gli elementi grafici a specifiche angolazioni e distanze dal centro del grafico. Nelle visualizzazioni 3-D è presente un'ulteriore dimensione di profondità.

Obliquo. Le trasformazioni oblique aggiungono un effetto 3-D agli elementi grafici. Viene applicata la profondità, che tuttavia è puramente decorativa e non influenzata da particolari valori dei dati.

Rapporto uguale. L'applicazione dello stesso rapporto consente di specificare che la stessa distanza in ciascuna scala rappresenta la stessa differenza tra i valori dei dati. Ad esempio, in entrambe le scale 2 cm rappresentano una differenza di 1000.

% riquadro pre-trasformazione. Se, dopo la trasformazione, gli assi vengono tagliati, prima di applicarla è possibile aggiungere riquadri nel grafico. I riquadri riducono di una determinata percentuale le dimensioni prima che vengano applicate le trasformazioni al sistema di coordinate. È possibile controllare le dimensioni x inferiore, x superiore, y inferiore e y superiore in quest'ordine.

% riquadro post-trasformazione. Se si desidera modificare le proporzioni del grafico, è possibile aggiungere riquadri al grafico dopo l'applicazione della trasformazione. I riquadri riducono di una determinata percentuale le dimensioni dopo l'applicazione di trasformazioni al sistema di coordinate. I riquadri possono essere applicati anche in assenza di trasformazioni. È possibile controllare le dimensioni x inferiore, x superiore, y inferiore e y superiore in quest'ordine.

Modifica di statistiche ed elementi grafici

È possibile convertire un elemento grafico in un altro tipo, modificare le statistiche utilizzate per tracciare l'elemento grafico o specificare il modificatore di collisione che determina l'operazione che viene eseguita quando gli elementi grafici si sovrappongono.

Come convertire un elemento grafico

1. Selezionare l'elemento grafico da convertire.
2. Fare clic sulla scheda **Elemento** nella palette delle proprietà.
3. Selezionare un nuovo tipo di elemento grafico dall'elenco Tipo.

Tabella 47. Tipi di elementi grafici

Tipo di elemento grafico	Descrizione
Point	Un marker che identifica un punto dati specifico. Un elemento punto viene utilizzato nei grafici a dispersione e in altre visualizzazioni correlate.
Interval	Una forma rettangolare disegnata in corrispondenza di un valore di dati specifico e che riempie lo spazio tra un'origine e un altro valore di dati. Un elemento intervallo viene utilizzato nei grafici a barre e negli istogrammi.
Line	Una linea che collega valori di dati.
Path	Una linea che collega valori di dati nell'ordine in cui appaiono nell'insieme di dati.
Area	Una linea che collega elementi di dati con l'area tra la linea e un'origine riempita.
Polygon	Una forma a più lati che racchiude un'area di dati. Un elemento poligono potrebbe essere usato in un grafico a dispersione in bin o in una mappa.
Schema	Un elemento composto da una scatola con baffi e simboli che indicano i valori anomali. Un elemento schema viene utilizzato per i grafici a scatole.

Come modificare le statistiche

1. Selezionare l'elemento grafico di cui si desidera modificare le statistiche.
2. Fare clic sulla scheda **Elemento** nella palette delle proprietà.
3. Dall'elenco a discesa Riepilogo selezionare una nuova statistica. Con la selezione di una statistica i dati vengono aggregati. Se, invece, si desidera visualizzare i dati separatamente, selezionare (**nessuna statistica**) dall'elenco Riepilogo.

Statistiche riassuntive calcolate a partire da un campo continuo

- *Media*. Una misura di tendenza centrale. Media aritmetica, ovvero somma divisa per il numero di casi.
- *Median*. È il valore sopra il quale e sotto il quale ricade la metà dei casi, il 50-esimo percentile. Se il numero di casi è pari, la mediana è pari alla media dei due casi centrali quando questi sono ordinati secondo l'ordine ascendente o discendente. La mediana è una misura di tendenza centrale non sensibile ai valori anomali, a differenza della media che può essere influenzata da valori eccezionalmente bassi o alti.
- *Modalità*. Il valore che ricorre più frequentemente. Se più valori condividono la maggiore ricorrenza, ognuno di essi è una modalità.
- *Minimo*. Il valore più basso di una variabile numerica.
- *Massimo*. Il valore più alto di una variabile numerica.
- *Intervallo*. Differenza tra i valori minimi e massimi.
- *Intervallo medio*. Il punto medio dell'intervallo, vale a dire, il valore la cui differenza rispetto al punto minimo è uguale alla differenza rispetto al punto massimo.
- *Somma*. La somma o il totale di tutti i valori non mancanti di tutti i casi.
- *Somma cumulata*. La somma cumulata dei valori. Ogni elemento grafico mostra la somma di un sottogruppo più la somma totale di tutti i gruppi precedenti.
- *Somma percentuale*. La percentuale all'interno di ogni sottogruppo basata su un campo sommato confrontato alla somma di tutti i gruppi.
- *Somma percentuale cumulata*. La percentuale cumulativa all'interno di ogni sottogruppo basata su un campo sommato confrontato alla somma di tutti i gruppi. Ogni elemento grafico mostra la percentuale di un sottogruppo più la percentuale totale di tutti i gruppi precedenti.
- *Varianza*. Una misura della dispersione dei valori intorno alla media. È calcolata come somma dei quadrati degli scostamenti dalla media, divisa per un valore corrispondente alla somma dei casi meno uno. La varianza è espressa in quadrati dell'unità di misura della variabile.
- *Deviazione standard*. Una misura di dispersione intorno alla media. In una distribuzione normale, il 68% dei casi rientra in una deviazione standard della media e il 95% dei casi rientra in due deviazioni standard. Se, ad esempio, in una popolazione con distribuzione normale l'età media fosse 45 e la deviazione standard 10, il 95% dei casi cadrebbe fra 25 e 65 anni.
- *Errore standard*. Una misura di quanto il valore di una statistica del test può variare da campione a campione. È la deviazione standard della distribuzione di campionamento di una statistica. L'errore standard della media, ad esempio, è la deviazione standard delle medie del campione.
- *Curtosi*. Una misura di quanto le osservazioni si trovino raggruppate in un punto centrale. Per la distribuzione normale, il valore della statistica di curtosi è zero. Una curtosi positiva indica che, rispetto a una distribuzione normale, le osservazioni sono più raggruppate intorno al centro della distribuzione e hanno code più sottili fino ai valori estremi della distribuzione; a quel punto, le code della distribuzione leptocurtica sono più spesse rispetto a una distribuzione normale. Una curtosi negativa indica che, rispetto a una distribuzione normale, le osservazioni sono meno raggruppate e hanno code più spesse fino ai valori estremi della distribuzione; a quel punto, le code della distribuzione platicurtica sono più sottili rispetto a una distribuzione normale.
- *Asimmetria*. Una misura dell'asimmetria di una distribuzione. La distribuzione normale è simmetrica e ha un valore di asimmetria pari a 0. Una distribuzione con una notevole asimmetria positiva ha una

lunga coda a destra. Una distribuzione con asimmetria negativa ha una coda a sinistra. In generale un'asimmetria con valore più che doppio dell'errore standard indica lo scostamento dalla normale simmetria.

Le seguenti statistiche di regione potrebbero restituire più elementi grafici per ogni sottogruppo. Quando si utilizzano gli elementi grafici intervallo, area o margine, una statistica di regione restituisce un elemento grafico che mostra l'intervallo. Tutti gli altri elementi grafici danno come risultato due elementi separati, uno che mostra l'inizio dell'intervallo e l'altro che mostra la fine dell'intervallo.

- **Regione: Intervallo.** L'intervallo di valori tra il valore minimo e il valore massimo.
- **Regione: Intervallo di confidenza per la media al 95%.** Un intervallo di valori con il 95% di probabilità di includere la media della popolazione.
- **Regione: Intervallo di confidenza per il singolo al 95% .** Un intervallo di valori con il 95% di probabilità di includere il valore previsto dato il caso singolo.
- **Regione: 1 Deviazione standard sopra/sotto la media.** Un intervallo di valori compreso tra 1 *deviazione standard* sopra e sotto la *media*.
- **Regione: 1 Errore standard sopra/sotto la media.** Un intervallo di valori compreso tra 1 *errore standard* sopra e sotto la *media*.

Statistiche riassuntive basate sul conteggio

- **Conteggio.** Il numero di righe/casi.
- **Conteggio cumulato.** Il numero cumulato di righe/casi. Ogni elemento grafico mostra il conteggio riferito a un sottogruppo più il conteggio totale di tutti i gruppi precedenti.
- **Percentuale del conteggio.** La percentuale di righe/casi in ogni sottogruppo rispetto al numero totale di righe/casi.
- **Percentuale cumulata del conteggio.** La percentuale cumulata di righe/casi in ogni sottogruppo rispetto al numero totale di righe/casi. Ogni elemento grafico mostra la percentuale di un sottogruppo più la percentuale totale di tutti i gruppi precedenti.

Come specificare il modificatore di collisione

Il modificatore di collisione determina cosa accade in caso di sovrapposizione tra elementi grafici.

1. Selezionare l'elemento grafico per il quale si desidera specificare il modificatore di collisione.
2. Fare clic sulla scheda **Elemento** nella palette delle proprietà.
3. Dall'elenco a discesa Modificatore selezionare un modificatore di collisione. Con l'impostazione **-automatico-**, è l'applicazione a stabilire il modificatore di collisione adatto al tipo e alle statistiche dell'elemento grafico.

Sovrapposto. Consente di disegnare elementi grafici uno sopra l'altro quando hanno lo stesso valore.

Sovrapposti. Consente di impilare elementi grafici che sarebbero normalmente sovrapposti quando presentano gli stessi valori dei dati.

Alternati. Consente di spostare gli elementi grafici accanto ad altri con lo stesso valore anziché sovrapporli. Gli elementi grafici vengono disposti in modo simmetrico. In altre parole, vengono spostati dalla parte opposta rispetto a un punto centrale. L'alternanza è molto simile al raggruppamento.

Impila. Consente di spostare gli elementi grafici accanto ad altri con lo stesso valore anziché sovrapporli. Gli elementi grafici vengono disposti in modo asimmetrico. In altri termini, gli elementi grafici vengono impilati uno sopra l'altro, con quello in basso posizionato in corrispondenza di uno specifico valore sulla scala.

Jitter (normale). Consente di riposizionare in modo casuale gli elementi grafici in corrispondenza di uno stesso valore dei dati utilizzando una distribuzione normale.

Jitter (uniforme). Consente di riposizionare in modo casuale gli elementi grafici in corrispondenza di uno stesso valore dei dati utilizzando una distribuzione uniforme.

Modifica della posizione della legenda

Se nel grafico è presente una legenda, quest'ultima di solito viene visualizzata sulla destra. Se necessario, tale posizione può essere modificata.

Come modificare la posizione della legenda

1. Selezionare la legenda.
2. Fare clic sulla scheda **Legenda** nella palette delle proprietà.
3. Selezionare una posizione.

Copia di una visualizzazione e dei dati delle visualizzazioni

La palette Generale comprende i pulsanti per la copia della visualizzazione e dei relativi dati.



Figura 64. Pulsante Copia visualizzazione

Copia visualizzazione. La visualizzazione viene copiata negli Appunti come immagine (sono disponibili più formati). Quando si incolla l'immagine in un'altra applicazione, è possibile selezionare un'opzione di "incolla" speciale che consente di scegliere il formato da incollare.



Figura 65. Pulsante Copia dati visualizzazione

Copia dati visualizzazione. Vengono copiati i dati sottostanti utilizzati per generare la visualizzazione. I dati vengono copiati negli Appunti come solo testo o testo HTML. Quando si incollano i dati in un'altra applicazione, è possibile selezionare un'opzione di "incolla" speciale che consente di scegliere il formato da incollare.

Tasti di scelta rapida

Tabella 48. Tasti di scelta rapida

Tasto di scelta rapida	Funzione
Ctrl+Barra spaziatrice	Alterna l'attivazione della modalità Esplora e Modifica
Canc	Consente di eliminare un oggetto della visualizzazione
Ctrl+Z	Annulla
Ctrl+Y	Ripeti
F2	Visualizza la struttura per selezionare gli elementi di un grafico

Aggiunta di titoli e annotazioni

È possibile aggiungere titoli, note a piè di pagina o etichette dell'asse a qualsiasi tipo di grafico in modo da facilitare l'identificazione dei dati visualizzati.

Aggiunta di titoli ai grafici

1. Dai menu, scegliere **Modifica > Aggiungi titolo grafico**. Sopra il grafico viene visualizzata una casella di testo contenente <TITOLO>.

2. Controllare di essere in modalità di modifica. Dai menu, scegliere **Visualizza > Modalità Modifica**.
3. Fare doppio clic sul testo <TITOLO>.
4. Immettere il titolo richiesto e premere Invio.

Aggiunta di note a piè di pagina ai grafici

1. Dai menu, scegliere **Modifica> Aggiungi piè di pagina grafico**. Sotto il grafico viene visualizzata una casella di testo contenente <PIÈ DI PAGINA>.
2. Controllare di essere in modalità di modifica. Dai menu, scegliere **Visualizza > Modalità Modifica**.
3. Fare doppio clic sul testo <PIÈ DI PAGINA>.
4. Immettere il titolo richiesto e premere Invio.

Utilizzo dei fogli di stile dei grafici

Le informazioni di base sulla visualizzazione dei grafici, quali colori, caratteri, simboli e spessore delle righe sono controllate da un foglio di stile. Con IBM SPSS Modeler viene fornito un foglio di stile di default; tuttavia, se necessario, è possibile modificarlo. Per esempio, se esiste uno schema di colori aziendale specifico per le presentazioni, potrebbe essere necessario utilizzarlo anche nei grafici. Per ulteriori informazioni, consultare l'argomento "Modifica delle visualizzazioni" a pagina 254.

Nei nodi Grafico è possibile utilizzare la modalità di modifica per cambiare lo stile dell'aspetto di un grafico. Tramite il menu **Modifica > Stili** è possibile quindi salvare le modifiche come un foglio di stile da applicare a tutti i grafici generati in seguito dal nodo Grafico corrente o da adottare come nuovo foglio di stile di default per tutti i grafici generati con IBM SPSS Modeler.

Dall'opzione **Stili** del menu Modifica sono disponibili cinque diverse opzioni per i fogli di stile:

- **Alterna foglio di stile.** Visualizza un elenco di diversi fogli di stile archiviati che possono essere selezionati per modificare l'aspetto dei grafici. Per ulteriori informazioni, consultare l'argomento "Applicazione di fogli di stile".
- **Archivia stili nel nodo.** Consente di archiviare le modifiche apportate agli stili del grafico selezionato in modo che vengano applicate a tutti i grafici creati successivamente dallo stesso nodo Grafico nel flusso corrente.
- **Archivia stili come stili di default.** Consente di archiviare le modifiche apportate agli stili del grafico selezionato in modo che vengano applicate a tutti i grafici creati successivamente da qualsiasi nodo Grafico in qualsiasi stream. Dopo aver selezionato questa opzione, è possibile utilizzare **Applica stili di default** per applicare gli stessi stili a tutti gli altri grafici esistenti.
- **Applica stili di default.** Consente di modificare gli stili del grafico selezionato applicando gli stili attualmente salvati come stili di default.
- **Applica stili originali.** Consente di ripristinare gli stili del grafico forniti come stili di default originali.

Applicazione di fogli di stile

È possibile applicare un foglio di stile di visualizzazione che specifichi le proprietà stilistiche della visualizzazione. Fra le altre opzioni, il foglio di stile può definire i caratteri, i tratteggi e i colori. In un certo senso, i fogli di stile rappresentano una "scorciatoia" per le modifiche che sarebbe necessario eseguire manualmente. Si noti, tuttavia, che un foglio di stile consente di eseguire solo le modifiche di *stile*. Le modifiche di altro tipo, come la posizione della legenda o l'intervallo della scala, non vengono memorizzate nel foglio di stile.

Come applicare un foglio di stile

1. Dai menu, scegliere:
Modifica > Stili > Alterna foglio di stile
2. Per selezionare un foglio di stile, utilizzare la finestra di dialogo Alterna foglio di stile.
3. Fare clic su **Applica** per applicare il foglio di stile alla visualizzazione senza chiudere la finestra. Fare clic su **OK** per applicare il foglio di stile e chiudere la finestra di dialogo.

Finestra di dialogo Alterna/Seleziona foglio di stile

La tabella nella parte superiore della finestra di dialogo elenca tutti i fogli di stile di visualizzazione attualmente disponibili. Alcuni fogli di stile sono preinstallati, mentre altri potrebbero essere stati creati in IBM SPSS Visualization Designer (un prodotto separato).

La parte inferiore della finestra di dialogo mostra alcuni esempi di visualizzazione contenenti dei dati campione. Selezionare uno dei fogli di stile in modo da applicare gli stili alle visualizzazioni di esempio. Questi esempi consentono di capire in che modo un determinato foglio di stile influisce effettivamente sulla visualizzazione.

Nella finestra di dialogo sono presenti anche le seguenti opzioni.

Stili esistenti. Per default, un foglio di stile può sovrascrivere tutti gli stili nella visualizzazione. Questo comportamento può essere modificato.

- **Sovrascrivi tutti gli stili.** Quando si applica il foglio di stile, consente di sovrascrivere tutti gli stili della visualizzazione, inclusi quelli modificati nella visualizzazione durante la sessione di modifica corrente.
- **Mantieni stili modificati.** Quando si applica il foglio di stile, consente di sovrascrivere solo gli stili che *non* sono stati modificati nella visualizzazione durante la sessione di modifica corrente. Al contrario, gli stili che sono stati modificati durante la sessione corrente vengono mantenuti.

Gestisci. Gestire modelli, fogli di stile e mappe di visualizzazione sul computer. È possibile importare, esportare, rinominare ed eliminare modelli, fogli di stile e mappe di visualizzazione sul computer locale. Per ulteriori informazioni, consultare l'argomento "Gestione dei modelli, dei fogli di stile e dei file delle mappe" a pagina 196.

Posizione. Modificare la posizione in cui sono archiviati i modelli, i fogli di stile e le mappe di visualizzazione. La posizione corrente viene elencata alla destra del pulsante. Per ulteriori informazioni, consultare l'argomento "Impostazione della posizione di modelli, fogli di stile e mappe" a pagina 196.

Stampa, salvataggio, copia ed esportazione di grafici

Ciascun grafico dispone di opzioni che consentono di salvare o stampare il grafico o di esportarlo in un altro formato. La maggior parte di queste opzioni è accessibile dal menu File. Inoltre, nel menu Modifica è possibile scegliere di copiare il grafico o i dati contenuti al suo interno in modo da poterli utilizzare in un'altra applicazione.

Stampa

Per stampare il grafico, utilizzare l'opzione di menu o il pulsante **Stampa**. Prima di stampare è possibile impostare le opzioni di stampa e visualizzare un'anteprima del grafico da stampare mediante i comandi **Imposta pagina** e **Anteprima di stampa**. Per ulteriori informazioni relative alla configurazione di intestazioni e piè di pagina, consultare .

Salvataggio di grafici

Per salvare il grafico in un file di output di IBM SPSS Modeler (*.cou), selezionare **File > Salva** oppure **File > Salva con nome** dai menu.

o

Per salvare il grafico nel repository, selezionare **File > Memorizza Output** dai menu.

Copia di grafici

Per copiare il grafico ed utilizzarlo in un'altra applicazione, come, ad esempio, MS Word o MS PowerPoint, selezionare **Modifica > Copia grafico** dai menu.

Copia di dati

Per copiare i dati in modo da utilizzarli in un'altra applicazione, per esempio MS Excel o MS Word, scegliere **Modifica > Copia dati del grafico**. Per default, i dati saranno copiati in formato HTML. Utilizzare **Incolla speciale...** nell'altra applicazione per visualizzare le altre opzioni di formattazione quando si incollano i dati.

Esportazione di grafici

L'opzione **Esporta grafico** consente di esportare il grafico in uno dei seguenti formati: bitmap (*.bmp*), JPEG (*.jpg*), PNG (*.png*), HTML (*.html*) o documento ViZml (*.xml*) per utilizzarlo in altre applicazioni IBM SPSS Statistics.

Per esportare i grafici, selezionare **File > Esporta grafico** e quindi scegliere il formato desiderato.

Esportazione di tabelle

L'opzione **Esporta tabella** consente di esportare la tabella in uno di seguenti formati: delimitato da tabulazione (*.tab*), delimitato da virgola (*.csv*) o HTML (*.html*).

Per esportare le tabelle, selezionare **File > Esporta tabella** dai menu, quindi scegliere il formato.

Capitolo 6. Nodi Output

Panoramica sui nodi Output

I nodi Output consentono di ottenere informazioni sui dati e sui modelli. Inoltre, consentono di esportare i dati in vari formati per l'interfacciamento con gli altri strumenti software in uso.

Sono disponibili i seguenti nodi di output:



Il nodo Tabella visualizza i dati in formato tabella, che è inoltre possibile scrivere su un file. Questa funzione è utile tutte le volte che si desidera controllare i valori dei dati o esportarli in un formato di facile lettura.



Il nodo Matrice crea una tabella che mostra le relazioni tra i campi. In genere viene utilizzato per mostrare le relazioni tra due campi simbolici, ma è possibile avvalersene anche per mostrare le relazioni tra campi flag o numerici.



Il nodo Analisi valuta la capacità dei modelli predittivi di generare previsioni accurate. I nodi Analisi eseguono diversi confronti tra i valori previsti e i valori effettivi per uno o più nugget del modello. Possono inoltre confrontare i modelli predittivi fra loro.



Il nodo Esplora offre una prima panoramica completa dei dati, incluse statistiche riassuntive, istogrammi e distribuzione per ciascun campo, nonché informazioni su valori anomali, mancanti ed estremi. I risultati vengono visualizzati in una matrice di semplice lettura che può essere ordinata e utilizzata per generare grafici a schermo intero e nodi di preparazione dei dati.



Il nodo Trasformazioni consente di selezionare e visualizzare in anteprima i risultati di trasformazioni prima di applicarli ai campi selezionati.



Il nodo Statistiche fornisce informazioni riassuntive di base su campi numerici. Calcola statistiche riassuntive per singoli campi e per correlazioni tra campi.



Il nodo Medie confronta le medie tra gruppi indipendenti o coppie di campi correlati per verificare se esiste una differenza significativa. Per esempio, è possibile confrontare le entrate medie prima e dopo il lancio di una promozione, oppure confrontare le entrate determinate da clienti che non hanno ricevuto la promozione con quelli che l'hanno ricevuta.



Il nodo Report crea report formattati che contengono sia testo fisso sia dati e altre espressioni derivate dai dati. Il formato del report viene specificato utilizzando modelli di testo per definire il testo fisso e costruzioni di output dei dati. È possibile fornire una formattazione personalizzata del testo utilizzando tag HTML nel modello e impostando apposite opzioni nella scheda Output. È possibile includere valori di dati e altro output condizionale utilizzando espressioni CLEM nel modello.



Il nodo Calcola globali analizza i dati e calcola i valori di riepilogo che possono essere utilizzati nelle espressioni CLEM. Per esempio, è possibile utilizzare questo nodo per calcolare le statistiche di un campo denominato *età* e utilizzare quindi la media globale dell'*età* nelle espressioni CLEM inserendo la funzione @GLOBAL_MEAN(*età*).



Il nodo Adattamento simulazione esamina la distribuzione statistica dei dati in ciascun campo e genera (o aggiorna) un nodo Genera simulazione con la distribuzione più adatta assegnata a ciascun campo. Il nodo Genera simulazione può essere quindi utilizzato per generare dati simulati.



Il nodo Valutazione simulazione valuta un campo obiettivo previsto specificato e presenta le informazioni della distribuzione e della correlazione relative al campo obiettivo.

Gestione dell'output

La scheda Output di Manager mostra i diagrammi, i grafici e le tabelle generati durante una sessione di IBM SPSS Modeler. È sempre possibile riaprire un output facendo doppio clic su di esso in Manager — non è necessario eseguire nuovamente il nodo o il flusso corrispondenti.

Per visualizzare la scheda Output di Manager

Aprire il menu Visualizza e scegliere **Manager**. Fare clic sulla scheda **Output**.

Dalla scheda Output di Manager è possibile:

- Visualizzare gli oggetti di output esistenti, per esempio istogrammi, grafici di valutazione e tabelle.
- Rinominare gli oggetti di output.
- Salvare gli oggetti di output su disco o, se disponibile, in IBM SPSS Collaboration and Deployment Services Repository.
- Aggiungere file di output al progetto corrente.
- Eliminare dalla sessione corrente gli oggetti di output non salvati.
- Aprire gli oggetti di output salvati o recuperarli da IBM SPSS Collaboration and Deployment Services Repository (se disponibile).

Per accedere a queste opzioni, fare clic con il pulsante destro del mouse in un punto qualsiasi della scheda Output.

Visualizzazione dell'output

L'output su schermo viene visualizzato in una finestra di output del browser. Questa finestra dispone dei propri menu che consentono di stampare e di salvare l'output o di esportarlo in un altro formato. Si noti che le opzioni specifiche possono variare a seconda del tipo di output.

Stampa, salvataggio ed esportazione di dati. Sono disponibili ulteriori informazioni, come segue:

- Per stampare l'output, utilizzare l'opzione di menu o il pulsante **Stampa**. Prima di stampare è possibile impostare le opzioni di stampa e visualizzare un'anteprima del grafico da stampare mediante i comandi **Imposta pagina** e **Anteprima di stampa**.
- Per salvare l'output in un file di output di IBM SPSS Modeler (*.cou*), scegliere **Salva** o **Salva con nome** dal menu File.

- Per salvare l'output in un altro formato, per esempio testo o HTML, scegliere **Esporta** dal menu File. Per ulteriori informazioni, consultare l'argomento "Esportazione dell'output" a pagina 274.
- Per salvare l'output in un repository condiviso e consentire ad altri utenti di visualizzarlo mediante IBM SPSS Collaboration and Deployment Services Deployment Portal, scegliere **Pubblica su Web** dal menu File. Questa opzione richiede una licenza distinta per IBM SPSS Collaboration and Deployment Services.

Selezione di celle e di colonne. Il menu Modifica contiene diverse opzioni per la selezione, la deselegione e la copia di celle e colonne, a seconda di quanto richiesto dal tipo di output corrente. Per ulteriori informazioni, consultare l'argomento "Selezione di celle e di colonne" a pagina 275.

Generazione di nodi. Il menu Genera consente di generare nuovi nodi in base al contenuto del browser di output. Le opzioni variano in base al tipo di output e agli elementi attualmente selezionati nell'output. Per informazioni sulle opzioni di generazione dei nodi di un tipo specifico di output, consultare la documentazione relativa al quel tipo di output.

Pubblica sul Web

L'opzione Pubblica su Web permette di pubblicare alcuni tipi di output dei flussi in un IBM SPSS Collaboration and Deployment Services Repository centralizzato condiviso che costituisce la base di IBM SPSS Collaboration and Deployment Services. Se si utilizza questa opzione, gli altri utenti che desiderano visualizzare l'output possono farlo se dispongono di un accesso a Internet e di un account IBM SPSS Collaboration and Deployment Services. Non è necessario che abbiano installato IBM SPSS Modeler.

Nota: per l'accesso ad un repository IBM SPSS Collaboration and Deployment Services è richiesta una licenza separata. Per ulteriori informazioni, consultare <http://www.ibm.com/software/analytics/spss/products/deployment/cds/>

La tabella riportata di seguito elenca i nodi IBM SPSS Modeler che supportano la funzione Pubblica su Web. L'output di questi nodi viene archiviato nel IBM SPSS Collaboration and Deployment Services Repository in formato oggetto di output (.cou) e può essere visualizzato direttamente in IBM SPSS Collaboration and Deployment Services Deployment Portal.

Gli altri tipi di output si possono visualizzare solo se sul computer dell'utente è installata la rispettiva applicazione (per esempio IBM SPSS Modeler per gli oggetti flusso).

Tabella 49. Nodi che supportano la funzione Pubblica su Web

Tipo di nodo	Nodo
Grafici	all
Output	Tabella
	Matrice
	Comandi di manipolazione dei nodi
	Trasforma
	Medie
	Analisi
	Statistiche
	Report (HTML)
IBM SPSS Statistics	Output Statistics

Pubblicazione dell'output sul Web

Per pubblicare l'output su Web:

1. Eseguire uno dei nodi elencati nella tabella in un flusso di IBM SPSS Modeler. Viene così creato un oggetto di output (per esempio un oggetto tabella, matrice o report) in una nuova finestra.
2. Dalla finestra dell'oggetto di output scegliere:
File > Pubblica sul Web
*Nota: se si desidera esportare semplici file HTML da utilizzare con un browser Web standard, selezionare **Esporta** dal menu File e selezionare **HTML**.*
3. Connessione a IBM SPSS Collaboration and Deployment Services Repository
Una volta effettuata correttamente la connessione, viene visualizzata la finestra di dialogo Repository: Archivia, che contiene diverse opzioni di archiviazione.
4. Dopo aver scelto le opzioni di archiviazione desiderate, fare clic su **Archivia**.

Visualizzazione dell'output pubblicato sul Web

Per usufruire di questa funzione è necessario disporre di un account IBM SPSS Collaboration and Deployment Services. Se si dispone dell'applicazione appropriata per visualizzare il tipo di oggetto desiderato (per esempio IBM SPSS Modeler o IBM SPSS Statistics), l'output viene visualizzato all'interno dell'applicazione anziché nel browser.

Nota: per l'accesso a IBM SPSS Collaboration and Deployment Services è necessaria una licenza separata. Per ulteriori informazioni, consultare <http://www.ibm.com/software/analytics/spss/products/deployment/cds/>.

Per visualizzare l'output pubblicato sul Web:

1. Dal browser, accedere a *http://<host_repos>:<port_repos>/peb*
dove *host_repos* e *port_repos* sono il nome host e il numero di porta dell'host IBM SPSS Collaboration and Deployment Services.
2. Digitare le credenziali di accesso per l'account IBM SPSS Collaboration and Deployment Services.
3. Fare clic su **Repository contenuti**.
4. Spostarsi sull'oggetto da visualizzare o cercarlo.
5. Fare clic sul nome dell'oggetto. Per alcuni tipi di oggetti, per esempio i grafici, la visualizzazione nel browser può non essere immediata.

Visualizzazione dell'output in un browser HTML

Nella scheda Opzioni avanzate dei nugget del modello Lineare, Logistico e Fattoriale/PCA è possibile visualizzare le informazioni in un browser separato, per esempio Internet Explorer. Le informazioni vengono prodotte in formato HTML, il che consente di salvarle e riutilizzarle altrove, per esempio in un'intranet aziendale o in un sito Internet.

Per visualizzare le informazioni in un browser, fare clic sul pulsante di avvio sotto l'icona del modello nell'angolo superiore sinistro della scheda Opzioni avanzate del nugget del modello.

Esportazione dell'output

Nella finestra di output del browser è possibile scegliere di esportare l'output in un altro formato, come testo o HTML. I formati di esportazione variano a seconda del tipo di output, ma solitamente sono simili alle opzioni relative al tipo di file disponibili quando si seleziona **Salva nel file** nel nodo utilizzato per generare l'output.

Per esportare l'output

1. Nel browser di output, aprire il menu File e scegliere **Esporta**. Quindi, scegliere il tipo di file che si desidera creare:

- **Delimitato da tabulazione (*.tab)** Questa opzione genera un file di testo formattato che contiene i valori dei dati. Questo stile spesso è utile per generare una rappresentazione in testo normale delle informazioni che possono essere importate in altre applicazioni. Questa opzione è disponibile per i nodi Tabella, Matrice e Medie.
- **Delimitato da virgola (*.dat)**. Questa opzione genera un file di testo delimitato da virgola contenente i valori dei dati. Questo stile è spesso utile per generare rapidamente un file di dati che può essere importato in fogli di calcolo o in altre applicazioni di analisi dei dati. Questa opzione è disponibile per i nodi Tabella, Matrice e Medie.
- **Dati trasposti delimitati da tabulazione (*.tab)** Questa opzione è identica all'opzione Delimitato da tabulazione, tranne per il fatto che i dati vengono trasposti in modo che le righe rappresentino i campi e le colonne rappresentino i record.
- **Dati trasposti delimitati da virgola (*.dat)**. Questa opzione è identica all'opzione Delimitato da virgola, tranne per il fatto che i dati vengono trasposti in modo che le righe rappresentino i campi e le colonne rappresentino i record.
- **HTML (*.html)**. Questa opzione scrive in uno o più file l'output con formattazione HTML.

Selezione di celle e di colonne

Alcuni nodi, come i nodi Tabella, Matrice e Medie, generano output tabulare. Queste tabelle di output possono essere visualizzate e manipolate in modi simili, che comprendono la selezione di celle, la copia di una tabella o di una parte di essa negli Appunti, la generazione di nuovi nodi basati sulla selezione corrente e il salvataggio e la stampa della tabella.

Selezione di celle. Per selezionare una cella, fare clic su di essa. Per selezionare le celle incluse in un'area rettangolare, fare clic su un angolo dell'area desiderata, trascinare la selezione sull'angolo opposto dell'area e quindi rilasciare il pulsante del mouse. Per selezionare un'intera colonna, fare clic sulla relativa intestazione. Per selezionare più colonne, utilizzare il metodo Maiusc-clic o Ctrl-clic sulle relative intestazioni.

Quando si esegue una nuova selezione, la selezione precedente verrà cancellata. Per aggiungere una nuova selezione a una selezione esistente evitando di cancellare la selezione precedente, è sufficiente eseguire la selezione tenendo premuto il tasto Ctrl. È possibile utilizzare questo metodo per selezionare più aree non contigue della tabella. Il menu Modifica contiene anche le opzioni **Seleziona tutto** e **Cancella selezione**.

Riordinamento delle colonne. I browser di output dei nodi Tabella e Medie consentono di spostare le colonne nella tabella facendo clic sull'intestazione delle colonne e trascinandole nelle posizioni desiderate. È possibile spostare una sola colonna per volta.

Nodo Tabella

Il nodo Tabella crea una tabella contenente i valori dei dati in esame. Sono compresi tutti i campi e tutti i valori del flusso, per agevolare l'analisi dei valori dei dati o la loro esportazione in un formato facilmente leggibile. Se lo si desidera, è possibile evidenziare i record che soddisfano una determinata condizione.

Scheda Impostazioni del nodo Tabella

Evidenzia record in cui. Per evidenziare i record presenti nella tabella, immettere un'espressione CLEM che sia vera per tutti i record da evidenziare. Questa opzione è attivata solo se è selezionato **Output su schermo**.

Scheda Formato del nodo Tabella

La scheda Formato contiene le opzioni di definizione della formattazione per ognuno dei campi ed è condivisa con il nodo Tipo. Per ulteriori informazioni, consultare l'argomento "Scheda per l'impostazione dei formati dei campi" a pagina 120.

Scheda Output del nodo Output

Per i nodi che generano output in stile tabella, la scheda Output consente di specificare il formato e la posizione dei risultati.

Nome output. Specifica il nome dell'output generato durante l'esecuzione del nodo. **Automatico** sceglie un nome in base al nodo che genera l'output. Se necessario, è possibile selezionare **Personalizzato** per specificare un nome diverso.

Output su schermo (default). Crea un nuovo oggetto di output da visualizzare in linea. L'oggetto di output verrà visualizzato nella scheda Output della finestra dei manager dopo l'esecuzione del nodo di output.

Output su file. Salva l'output in un file dopo l'esecuzione del nodo. Se si sceglie questa opzione, immettere un nome di file (oppure passare a una directory e specificare un nome di file utilizzando il pulsante di selezione dei file) e selezionare un tipo di file. Si noti che alcuni tipi di file potrebbero non essere disponibili per determinati tipi di output.

I dati vengono prodotti nel formato di codifica di default del sistema, specificato nel Pannello di controllo di Windows oppure, se si esegue il programma in modalità distribuita, sul computer server.

- **Dati (delimitati da tabulazione) (*.tab).** Questa opzione genera un file di testo formattato che contiene i valori dei dati. Questo stile spesso è utile per generare una rappresentazione in testo normale delle informazioni che possono essere importate in altre applicazioni. Questa opzione è disponibile per i nodi Tabella, Matrice e Medie.
- **Dati (delimitati da virgole) (*.dat).** Questa opzione genera un file di testo delimitato da virgola contenente i valori dei dati. Questo stile è spesso utile per generare rapidamente un file di dati che può essere importato in fogli di calcolo o in altre applicazioni di analisi dei dati. Questa opzione è disponibile per i nodi Tabella, Matrice e Medie.
- **HTML (*.html).** Questa opzione scrive in uno o più file l'output con formattazione HTML. Nel caso di output tabulare (dai nodi Tabella, Matrice o Medie), un insieme di file HTML include un riquadro contenente l'elenco dei nomi dei campi e i dati in una tabella HTML. Se il numero di righe presenti nella tabella è superiore alla specifica **Righe per pagina**, la tabella può essere suddivisa in più file HTML. In questo caso, il riquadro del contenuto contiene i collegamenti a tutte le pagine della tabella e offre uno strumento per l'esplorazione della tabella. Se l'output non è di tipo tabulare, viene creato un solo file HTML contenente i risultati del nodo.

Nota: se l'output HTML contiene formattazione solo per la prima pagina, selezionare **Impagina output** e modificare la specifica **Righe per pagina** in modo da includere tutto l'output in una singola pagina. Se invece il modello di output per nodi quali il nodo Report contiene tag HTML personalizzati, assicurarsi di specificare il tipo di formato **Personalizzato**.

- **File di testo (*.txt).** Questa opzione genera un file di testo contenente l'output. Questo stile è spesso utile per generare un output che può essere importato in altre applicazioni, per esempio elaboratori di testo o software per presentazioni. Questa opzione non è disponibile per alcuni nodi.
- **Oggetto di output (*.cou).** Gli oggetti di output salvati in questo formato possono essere aperti e visualizzati in IBM SPSS Modeler, aggiunti ai progetti, nonché pubblicati e registrati mediante IBM SPSS Collaboration and Deployment Services Repository.

Visualizzazione output. Per il nodo Medie è possibile specificare se per default deve essere visualizzato l'output semplice o avanzato. Si noti che quando si visualizza l'output generato è anche possibile passare da un tipo di visualizzazione all'altro. Per ulteriori informazioni, consultare l'argomento "Browser di output del nodo Medie" a pagina 297.

Formato. Per il nodo Report, è possibile scegliere se l'output deve essere formattato automaticamente oppure utilizzando la formattazione HTML inclusa nel modello. Per consentire la formattazione HTML del modello, selezionare **Personalizzato**.

Titolo. Per il nodo Report, è possibile specificare un titolo facoltativo che verrà visualizzato all'inizio dell'output del report.

Evidenzia testo inserito. Per il nodo Report, selezionare questa opzione per evidenziare il testo generato da espressioni CLEM nel modello Report. Per ulteriori informazioni, consultare l'argomento "Scheda Modello del nodo Report" a pagina 299. Questa opzione non è consigliata se si utilizza una formattazione di tipo **Personalizzato**.

Righe per pagina. Per il nodo Report, specificare il numero di righe da includere in ogni pagina durante la formattazione di tipo **Automatico** del report di output.

Trasponi dati. Questa opzione traspone i dati prima dell'esportazione in modo che le righe rappresentino i campi e le colonne rappresentino i record.

Nota: se le tabelle sono di grandi dimensioni, le opzioni sopra riportate possono risultare non completamente efficienti, soprattutto quando si utilizza un server remoto. In questi casi, si ottengono prestazioni decisamente migliori utilizzando un nodo di output File. Per ulteriori informazioni, consultare l'argomento "Nodo di esportazione File flat" a pagina 331.

Browser del nodo Tabella

Il browser del nodo Tabella visualizza i dati in formato tabellare e consente di eseguire operazioni standard, tra cui la selezione e la copia di celle, il riordinamento delle colonne ed il salvataggio e la stampa della tabella. Per ulteriori informazioni, consultare l'argomento "Selezione di celle e di colonne" a pagina 275. Si tratta delle stesse operazioni che è possibile eseguire quando si visualizza l'anteprima dei dati di un nodo.

Esportazione dei dati delle tabelle. È possibile esportare i dati dal browser del nodo Tabella scegliendo:

File > Esporta

Per ulteriori informazioni, consultare l'argomento "Esportazione dell'output" a pagina 274.

I dati vengono esportati nel formato di codifica di default del sistema, specificato nel Pannello di controllo di Windows oppure, se si esegue il programma in modalità distribuita, sul computer server.

Ricerca nella tabella. Il pulsante di ricerca (indicato dall'icona binocolo) nella barra degli strumenti principale attiva la barra degli strumenti di ricerca, consentendo di ricercare valori specifici nella tabella. È possibile cercare in avanti o all'indietro nella tabella, specificare una ricerca con distinzione tra caratteri maiuscoli/minuscoli (pulsante **Aa**) e interrompere una ricerca in corso premendo il relativo pulsante di interruzione.

Generazione di nodi. Il menu Genera rende disponibili operazioni di generazione dei nodi.

- **Nodo Selezione ("Record").** Genera un nodo Selezione che seleziona i record per i quali è selezionata una qualsiasi cella nella tabella.
- **Nodo Selezione ("E").** Genera un nodo Selezione che seleziona i record che contengono *tutti* i valori selezionati nella tabella.
- **Nodo Selezione ("O").** Genera un nodo Selezione che seleziona i record che contengono *uno qualsiasi* dei valori selezionati nella tabella.
- **Ricava ("Record").** Genera un nodo Ricava per creare un nuovo campo flag. Il campo flag contiene *T* per i record per i quali è selezionata una qualsiasi cella nella tabella e *F* per i record restanti.
- **Ricava ("E").** Genera un nodo Ricava per creare un nuovo campo flag. Il campo flag contiene *T* per i record che contengono *tutti* i valori selezionati nella tabella e *F* per i record restanti.
- **Ricava ("O").** Genera un nodo Ricava per creare un nuovo campo flag. Il campo flag contiene *T* per i record che contengono *uno qualsiasi* dei valori selezionati nella tabella e *F* per i record restanti.

Nodo Matrice

Il nodo Matrice consente di creare una tabella che mostra le relazioni tra i campi. In genere viene utilizzato per mostrare le relazioni tra due campi categoriali (flag, nominali o ordinali), ma è possibile avvalersene anche per mostrare le relazioni tra campi continui (intervallo numerico).

Scheda Impostazioni del nodo Matrice

La scheda Impostazioni consente di specificare le opzioni per la struttura della matrice.

Campi. Selezionare una delle seguenti opzioni per specificare un tipo di selezione di campo:

- **Selezionati.** Questa opzione consente di selezionare un campo relativo alla categoria per le righe ed uno per le colonne della matrice. Le righe e le colonne della matrice sono definite dall'elenco dei valori per il campo categoriale selezionato. Le celle della matrice contengono le statistiche riassuntive selezionate sotto.
- **Tutti i flag (valori veri).** Questa opzione richiede una matrice con un'unica riga e un'unica colonna per ogni campo flag nei dati. Le celle della matrice contengono i conteggi dei positivi doppi per ogni combinazione di flag. In altre parole, per una riga che corrisponde a *pane acquistato* e una colonna che corrisponde a *formaggio acquistato*, la cella che si trova nel punto di intersezione tra tale riga e tale colonna contiene il numero di record per i quali sia *pane acquistato* sia *formaggio acquistato* sono veri.
- **Tutti i valori numerici.** Questa opzione richiede una matrice con un'unica riga e un'unica colonna per ogni campo numerico. Le celle della matrice rappresentano la somma dei prodotti incrociati per la coppia di campi corrispondente. In altre parole, per ogni cella presente nella matrice, i valori per il campo riga e i valori per il campo colonna vengono moltiplicati per ogni record e poi sommati tra i vari record.

Includi valori mancanti. Consente di includere i valori mancanti definiti dall'utente (vuoti) e i valori mancanti di sistema (\$null\$) nell'output delle righe e delle colonne. Per esempio, se il valore *N/A* è stato specificato come mancante definito dall'utente per il campo della colonna selezionato, nella tabella verrà inclusa una colonna separata denominata *N/A* (a condizione che questo valore ricorra effettivamente nei dati) proprio come per tutte le altre categorie. Se l'opzione non è selezionata, la colonna *N/A* verrà esclusa, indipendentemente da quante volte ricorre il valore.

Nota: l'opzione di inclusione dei valori mancanti è valida solo quando i campi selezionati sono tavole di contingenza. I valori vuoti vengono mappati a \$null\$ ed esclusi dall'aggregazione per il campo della funzione quando la modalità è **Selezionati** e il contenuto è impostato su **Funzione** e per tutti i campi numerici quando la modalità è impostata su **Tutti i valori numerici**.

Contenuto celle. Se sono stati scelti campi **Selezionati**, è possibile specificare la statistica da utilizzare nelle celle della matrice. Selezionare una statistica basata sul conteggio oppure un campo di sovrapposizione per riassumere i valori di un campo numerico in base ai valori dei campi riga e colonna.

- **Tavole di contingenza.** I valori delle celle sono conteggi e/o percentuali del numero di record che includono la combinazione di valori corrispondente. È possibile specificare i riepiloghi delle tavole di contingenza desiderati, utilizzando le opzioni della scheda **Aspetto**. Insieme alla significatività viene visualizzato anche il valore chi-quadrato globale. Per ulteriori informazioni, consultare l'argomento "Browser di output del nodo Matrice" a pagina 279.
- **Funzione.** Se si seleziona una funzione riassuntiva, i valori delle celle sono rappresentati da una funzione dei valori del campo di sovrapposizione per i casi che includono i valori di riga e di colonna appropriati. Per esempio, se il campo riga è *Regione*, il campo colonna è *Prodotto* e il campo di sovrapposizione è *Entrate*, la cella di intersezione tra la riga *Nordest* e la colonna *Gadget* conterrà la somma (la media, il valore minimo o il valore massimo) delle entrate per i gadget venduti nella regione nordest. La funzione riassuntiva di default è **Media**. È possibile selezionare un'altra funzione per riassumere il campo funzione. Le opzioni includono **Media**, **Somma**, **DevS** (deviazione standard), **Max** (massimo) e **Min** (minimo).

Scheda Aspetto del nodo Matrice

La scheda Aspetto consente di controllare le opzioni di ordinamento ed evidenziazione per la matrice e le statistiche visualizzate per le matrici delle tavole di contingenza.

Righe e colonne. Controlla l'ordinamento delle intestazioni di riga e colonna nella matrice. L'impostazione di default è **Non ordinate**. Selezionare **Crescente** o **Decrescente** per ordinare le intestazioni di riga e colonna nel modo specificato.

Sovrapposizione. Consente di evidenziare i valori estremi nella matrice. I valori sono evidenziati in base ai conteggi delle celle (per le matrici delle tavole di contingenza) o ai valori calcolati (per le matrici di funzione).

- **Evidenzia i primi.** È possibile richiedere che vengano evidenziati (in rosso) i valori massimi presenti nella matrice. Specificare il numero di valori da evidenziare.
- **Evidenzia gli ultimi.** È inoltre possibile richiedere che vengano evidenziati (in verde) i valori minimi presenti nella matrice. Specificare il numero di valori da evidenziare.

Nota: per le due opzioni di evidenziazione, i vincoli tra i numeri possono determinare l'evidenziazione di un numero di valori maggiore di quello richiesto. Per esempio, se per una matrice con sei zeri tra le celle si specifica **Evidenzia gli ultimi 5**, verranno evidenziati tutti e sei gli zeri.

Contenuto delle tavole di contingenza. Per le tavole di contingenza è possibile specificare le statistiche riassuntive incluse nella matrice per le matrici delle tavole di contingenza. Queste opzioni non sono disponibili quando nella scheda Impostazioni è selezionata l'opzione **Tutti i valori numerici** o **Funzione**.

- **Conteggi.** Le celle includono il numero di record contenenti il valore di riga corrispondente al valore di colonna. Questo è l'unico contenuto della cella di default.
- **Valori attesi.** Il valore previsto per il numero di record nella cella, presumendo che non esista alcuna relazione tra le righe e le colonne. I valori previsti sono basati sulla seguente formula:
$$p(\text{row value}) * p(\text{column value}) * \text{total number of records}$$
- **Residui.** La differenza tra i valori osservati e quelli previsti.
- **Percentuale della riga.** La percentuale di tutti i record contenenti il valore di riga corrispondente al valore di colonna. La somma delle percentuali è 100 all'interno delle righe.
- **Percentuale della colonna.** La percentuale di tutti i record contenenti il valore di colonna corrispondente al valore di riga. La somma delle percentuali è 100 all'interno delle colonne.
- **Percentuale del totale.** La percentuale di tutti i record contenenti la combinazione di valore di colonna e valore di riga. La somma delle percentuali è 100 nell'intera matrice.
- **Includi totali di righe e colonne.** Aggiunge alla matrice una riga e una colonna per i totali di righe e colonne.
- **Applica impostazioni.** (Solo browser di output) Consente di modificare l'aspetto dell'output del nodo Matrice senza che sia necessario chiudere e riaprire il browser di output. Apportare le modifiche in questa scheda del browser di output, fare clic su questo pulsante, quindi selezionare la scheda Matrice per visualizzare l'effetto dei cambiamenti apportati.

Browser di output del nodo Matrice

Il browser del nodo Matrice visualizza i dati delle tavole di contingenza e consente di eseguire operazioni sulla matrice, tra cui la selezione di celle, la copia di tutta o di una parte della matrice negli Appunti, la generazione di nuovi nodi in base alla selezione della matrice ed il salvataggio e la stampa della matrice. Il browser del nodo Matrice può essere utilizzato anche per visualizzare l'output da alcuni modelli, come i modelli Naive Bayes da Oracle.

I menu File e Modifica contengono le normali opzioni di stampa, salvataggio, esportazione dell'output, selezione e copia dei dati. Per ulteriori informazioni, consultare l'argomento "Visualizzazione dell'output" a pagina 272.

Chi-quadrato. Per la tavola di contingenza di due campi categoriali, sotto alla tabella viene visualizzato anche il chi-quadrato di Pearson globale. Questo test indica la probabilità che i due campi non siano correlati in base alla differenza tra i conteggi osservati e i conteggi previsti se non esistesse alcuna relazione. Per esempio, se non esistesse alcuna relazione tra soddisfazione dei clienti e ubicazione del negozio, sarebbe normale aspettarsi percentuali di soddisfazione simili per tutti i negozi. Se però i clienti di alcuni punti vendita fanno registrare costantemente percentuali superiori rispetto agli altri, è logico sospettare che non si tratti di una coincidenza. Maggiore è la differenza, minore è la probabilità che il risultato sia dovuto semplicemente a un errore casuale di campionamento.

- Il test del chi-quadrato indica la probabilità che i due campi non siano correlati, nel qual caso le eventuali differenze tra le frequenze osservate e quelle previste sono determinate unicamente dal caso. Se questa probabilità è estremamente ridotta — generalmente meno del 5%— la relazione tra i due campi è definita significativa.
- In presenza di una sola colonna o di una sola riga (un test del chi-quadrato monodirezionale) i gradi di libertà sono rappresentati dal numero delle celle meno uno. Per un chi-quadrato bidirezionale, il grado di libertà è il numero di righe meno uno moltiplicato per il numero di colonne meno uno.
- La statistica del chi-quadrato deve essere interpretata con prudenza se una qualsiasi delle frequenze di cella previste è inferiore a cinque.
- Il test del chi-quadrato è disponibile solo per una tavola di contingenza di due campi. Quando nella scheda Impostazioni è selezionata l'opzione **Tutti i flag** o **Tutti i valori numerici**, il test non viene visualizzato.

Menu Genera. Il menu Genera rende disponibili operazioni di generazione dei nodi. Queste operazioni sono disponibili solo per matrici delle tavole di contingenza ed è necessario che nella matrice sia selezionata almeno una cella.

- **Nodo Seleziona.** Genera un nodo Seleziona che seleziona i record corrispondenti a qualsiasi cella selezionata nella matrice.
- **Nodo Ricava (flag).** Genera un nodo Ricava per creare un nuovo campo flag. Il campo flag contiene V per i record che corrispondono a qualsiasi cella selezionata nella matrice e F per i record restanti.
- **Nodo Ricava (insieme).** Genera un nodo Ricava per creare un nuovo campo nominale. Il campo nominale contiene una sola categoria per ogni insieme contiguo di celle selezionate nella matrice.

nodo Analisi

Il nodo Analisi consente di valutare la capacità di un modello di generare previsioni accurate. I nodi Analisi eseguono diversi confronti tra i valori previsti e i valori effettivi (il campo obiettivo) per uno o più nugget del modello. I nodi Analisi possono essere inoltre utilizzati per confrontare modelli predittivi con altri modelli predittivi.

Quando si esegue un nodo Analisi, alla sezione Analisi nella scheda Riepilogo viene automaticamente aggiunto un riepilogo dei risultati dell'analisi per ogni nugget del modello nel flusso eseguito. I risultati dettagliati dell'analisi vengono visualizzati nella scheda Output della finestra dei manager, oppure possono essere scritti direttamente in un file.

Nota: poiché i nodi Analisi confrontano i valori previsti con i valori effettivi, sono utili solo con i modelli supervisionati (quelli che richiedono un campo obiettivo). Per i modelli non supervisionati come gli algoritmi di cluster non esistono risultati effettivi disponibili da utilizzare come base per il confronto.

Scheda Analisi del nodo Analisi

La scheda Analisi consente di specificare i dettagli dell'analisi.

Matrici di coincidenza (per obiettivi simbolici o categoriali). Mostra lo schema di corrispondenze tra ogni campo generato (previsto) e il campo obiettivo corrispondente per gli obiettivi categoriali (flag, nominali o ordinali). Viene visualizzata una tabella in cui le righe sono definite dai valori effettivi e le colonne dai valori predittivi, con il numero di record che presentano questo schema in ogni cella. Questa

visualizzazione è utile per identificare gli errori sistematici nella previsione. Se esistono più campi generati che fanno riferimento allo stesso campo di output ma sono stati prodotti da modelli diversi, vengono conteggiate tutte le corrispondenze e le non corrispondenze di questi campi e vengono visualizzati i totali. Per i casi in cui i campi corrispondono, viene visualizzato un altro insieme di statistiche corrette/errate.

Valutazione performance. Mostra le statistiche di valutazione delle prestazioni per i modelli con output categoriali. Questa statistica, riportata per ogni categoria del campo o dei campi di output, è una misura del contenuto medio (in bit) delle informazioni del modello per la previsione dei record che appartengono a quella categoria. La statistica tiene in considerazione la difficoltà del problema di classificazione; pertanto, le previsioni più accurate per le categorie rare conseguiranno un indice di valutazione delle prestazioni più alto che le previsioni accurate per le categorie comuni. Se per una categoria il modello consente di ottenere solo semplici ipotesi, l'indice di valutazione delle prestazioni per tale categoria sarà 0.

Metriche di valutazione (solo classificatori binari, AUC e Gini). Per i classificatori binari, questa opzione riporta le metriche di valutazione con coefficiente Gini e AUC (area under curve). Entrambe le metriche di valutazione vengono calcolate insieme per ogni modello binario. I valori delle metriche vengono riportati in una tabella nel browser di output dell'analisi.

La metrica di valutazione AUC viene calcolata come area in una curva ROC (receiver operator characteristic) ed è una rappresentazione scalare delle prestazioni previste di un classificatore. L'AUC è sempre compresa tra 0 e 1, dove un numero più alto rappresenta un classificatore migliore. Una curva ROC diagonale tra le coordinate (0,0) e (1,1) rappresenta un classificatore casuale ed ha un'AUC di 0.5. Quindi un classificatore realistico non avrà un'AUC inferiore a 0.5.

La metrica di valutazione con coefficiente Gini a volte viene utilizzata come metrica di valutazione alternativa all'AUC, e le due misurazioni sono strettamente correlate. Il coefficiente Gini viene calcolato come due volte l'area compresa tra la curva ROC e la diagonale o come $Gini = 2AUC - 1$. Il coefficiente Gini è sempre compreso tra 0 e 1, dove un numero più alto rappresenta un classificatore migliore. Il coefficiente Gini è negativo nell'improbabile caso che la curva ROC sia sotto la diagonale.

Cifre di confidenza (se disponibili). Nel caso di modelli che generano un campo di confidenza, questa opzione consente di ottenere statistiche sui valori di confidenza e la loro relazione con le previsioni. Esistono due impostazioni per questa opzione:

- **Soglia di correzione.** Consente di ottenere informazioni sul livello di confidenza oltre il quale la precisione corrisponderà alla percentuale specificata.
- **Migliora precisione.** Consente di ottenere informazioni sul livello di confidenza oltre il quale la precisione migliora in base al fattore specificato. Per esempio, se la precisione globale è 90% e questa opzione è impostata su 2,0 il valore riportato corrisponderà alla confidenza richiesta per una precisione del 95%.

Trova campi previsti/predittore mediante. Determina in che modo i campi previsti vengono fatti corrispondere al campo obiettivo originale.

- **Metadati del campo di output del modello.** Fa corrispondere i campi previsti all'obiettivo in base alle informazioni sui campi del modello, consentendo la corrispondenza anche se un campo previsto è stato rinominato. Alle informazioni sui campi del modello è possibile accedere anche per qualsiasi campo previsto dalla finestra di dialogo Valori mediante un nodo tipologia. Per ulteriori informazioni, consultare l'argomento "Utilizzo della finestra di dialogo Valori" a pagina 116.
- **Formato del nome campo.** Fa corrispondere i campi in base alla convenzione di denominazione. Per esempio, i valori previsti generati da un nugget del modello C5.0 per un obiettivo denominato *risposta* devono trovarsi in un campo denominato *\$C-risposta*.

Suddividi per partizione. Se si utilizza un campo partizione per suddividere i record in campioni di addestramento, di test e di convalida, selezionare questa opzione per visualizzare i risultati separatamente per ogni partizione. Per ulteriori informazioni, consultare l'argomento "Nodo Partizione" a pagina 146.

Nota: quando viene eseguita la suddivisione per partizione, i record con valori null nel campo della partizione vengono esclusi dall'analisi. Ciò non rappresenta un problema se si utilizza un nodo Partizione, poiché i nodi Partizione non generano valori null.

Analisi definita dall'utente. È possibile specificare che nella valutazione del modello o dei modelli venga utilizzato un calcolo di analisi personalizzato. Utilizzare le espressioni CLEM per specificare l'oggetto da calcolare per ogni record e come combinare i punteggi a livello di record in un punteggio globale. Utilizzare le funzioni @TARGET e @PREDICTED per fare riferimento rispettivamente al valore obiettivo (output effettivo) e al valore previsto.

- **Se.** Specificare un'espressione condizionale se si desidera utilizzare calcoli diversi in base a condizioni determinate.
- **Allora.** Specificare il calcolo se la condizione Se è vera.
- **Altrimenti.** Specificare il calcolo se la condizione Se è falsa.
- **Utilizza.** Selezionare una statistica per calcolare il punteggio globale dai punteggi individuali.

Analisi distinta per campi. Mostra i campi categoriali disponibili per l'esecuzione di analisi distinte. Oltre all'analisi globale, verrà riportata un'analisi separata per ogni categoria di ciascun campo specificato.

Browser di output del nodo Analisi

Il browser di output del nodo Analisi visualizza i risultati dell'esecuzione del nodo Analisi. Dal menu File sono disponibili le normali opzioni di salvataggio, esportazione e stampa. Per ulteriori informazioni, consultare l'argomento "Visualizzazione dell'output" a pagina 272.

Quando si visualizza per la prima volta l'output del nodo Statistiche, i risultati vengono espansi. Per nascondere i risultati, utilizzare il controllo dell'espansore di un elemento se si desidera comprimere solo i risultati di tale elemento oppure fare clic sul pulsante **Comprimi tutto** se si desidera nascondere tutti i risultati. Per visualizzare nuovamente i risultati dopo averli compressi, utilizzare il controllo dell'espansore a sinistra di un elemento per visualizzarne i risultati, oppure fare clic sul pulsante **Espandi tutto** per visualizzare tutti i risultati.

Risultati per campo di output. L'output del nodo Analisi contiene una sezione per ogni campo di output per il quale esiste un campo di previsione corrispondente creato da un modello generato.

Confronto. Nella sezione del campo di output esiste una sottosezione per ogni campo di previsione associato a tale campo di output. Nel caso di campi di output categoriali, il livello superiore di questa sezione contiene una tabella che mostra il numero e la percentuale di previsioni corrette ed errate e il numero complessivo di record nel flusso. Nel caso di campi di output numerici, questa sezione mostra le seguenti informazioni:

- **Numero min di errori.** Mostra il numero minimo di errori (differenza tra i valori osservati e quelli previsti).
- **Numero max di errori.** Mostra il numero massimo di errori.
- **Errore media.** Mostra il valore medio (media) degli errori in tutti i record. La media mostra se nel modello esiste una **distorsione** sistematica (una maggiore tendenza alla stima per eccesso rispetto a quella per difetto o viceversa).
- **Errore assoluto della media.** Mostra la media dei valori assoluti degli errori in tutti i record. Indica la grandezza media dell'errore, indipendentemente dalla direzione.
- **Deviazione standard.** Mostra la deviazione standard degli errori.

- **Correlazione lineare.** Mostra la correlazione lineare tra i valori previsti e quelli effettivi. Questa statistica varia tra -1.0 e 1.0 . Valori vicini a $+1.0$ indicano una forte associazione positiva: in questo caso, i valori alti previsti sono associati ai valori alti effettivi e i valori bassi previsti sono associati ai valori bassi effettivi. I valori vicini a -1.0 indicano una forte associazione negativa, in modo che i valori alti previsti vengono associati a valori bassi effettivi e viceversa. Valori vicini a $0,0$ indicano un'associazione debole: in questo caso, i valori previsti sono più o meno indipendenti dai valori effettivi. *Nota:* una voce vuota indica che in questo caso non è possibile calcolare tale correlazione lineare, perché i valori previsti o effettivi sono costanti.
- **Occorrenze.** Mostra il numero di record utilizzati nell'analisi.

Matrice coincidenza. Se nelle opzioni di analisi è stata richiesta una matrice di coincidenza per campi di output categoriali, tale matrice viene inclusa in questa sottosezione dell'output. Le righe rappresentano i valori effettivi osservati e le colonne rappresentano i valori previsti. La cella nella tabella indica il numero di record per ogni combinazione di valori previsti e valori effettivi.

Valutazione performance. Se nelle opzioni di analisi è stata richiesta una statistica di valutazione delle prestazioni per campi di output categoriali, i risultati di tale valutazione vengono visualizzati in questa sezione dell'output. Viene elencata ogni categoria di output con la relativa statistica di valutazione delle prestazioni.

Report Valori di confidenza. Se nelle opzioni di analisi sono stati richiesti valori di confidenza per campi di output categoriali, tali valori vengono visualizzati in questa sezione. Per i valori di confidenza del modello vengono riportate le seguenti statistiche:

- **Intervallo.** Mostra l'intervallo (valori minimi e valori massimi) dei valori di confidenza per i record dei dati nel flusso.
- **Media corretta.** Mostra la confidenza media per i record classificati correttamente.
- **Media errata.** Mostra la confidenza media per i record classificati non correttamente.
- **Sempre corretto sopra.** Mostra la soglia di confidenza oltre la quale le previsioni sono sempre corrette e la percentuale dei casi che soddisfano questo criterio.
- **Sempre errato sotto.** Mostra la soglia di confidenza al di sotto della quale le previsioni sono sempre errate e la percentuale dei casi che soddisfano questo criterio.
- **X% precisione al di sopra di.** Mostra il livello di confidenza al quale la precisione è X%. X è approssimativamente il valore specificato per **Soglia di correzione** nelle opzioni Analisi. Per alcuni modelli e insiemi di dati non è possibile scegliere un valore di confidenza che fornisca la soglia esatta specificata nelle opzioni (in genere a causa di cluster di casi simili con lo stesso valore di confidenza vicino alla soglia). La soglia visualizzata nel report corrisponde al valore più vicino al criterio di precisione specificato che può essere ottenuto con un'unica soglia del valore di confidenza.
- **X volte corrette al di sopra di.** Mostra il valore di confidenza con una precisione di X volte superiore rispetto all'intero dataset. X è il valore specificato per **Migliora precisione** nelle opzioni Analisi.

Accordo tra. Se nel flusso vengono inclusi due o più modelli generati che prevedono lo stesso campo di output, verranno visualizzate anche le statistiche sull'**accordo** tra le previsioni generate dai modelli. Le statistiche includono il numero e la percentuale di record per i quali le previsioni concordano (per campi di output categoriali) o le statistiche riassuntive degli errori (per campi di output continui). Per i campi categoriali le statistiche includono un'analisi delle previsioni rispetto ai valori effettivi per il sottoinsieme di record in cui i modelli concordano (generano lo stesso valore previsto).

Metriche di valutazione. Per i classificatori binari, se sono state richieste metriche di valutazione nelle opzioni dell'analisi, i valori delle metriche di valutazione con coefficiente AUC e Gini sono visualizzati in una tabella in questa sezione. La tabella ha una riga per ogni modello di classificatore binario. La tabella delle metriche di valutazione viene visualizzata per ogni campo di output, piuttosto che per ogni modello.

Nodo Esplora

Il nodo Esplora fornisce una prima analisi completa dei dati portati all'interno di IBM SPSS Modeler, presentati in una matrice di semplice lettura che può essere ordinata e utilizzata per generare grafici a schermo intero e svariati nodi di preparazione dati.

- La scheda Esplora visualizza un report che fornisce statistiche riassuntive, istogrammi e grafici di distribuzione che possono essere utili per una comprensione iniziale dei dati. Nel report, inoltre, i nomi dei campi sono preceduti dall'icona di archiviazione.
- La scheda Qualità del report di verifica visualizza informazioni sui valori anomali, estremi e mancanti e offre gli strumenti necessari per gestire questi valori.

Utilizzo del nodo Esplora

Il nodo Esplora può essere collegato direttamente a un nodo origine oppure collocato a valle di un nodo Tipo istanziato. Inoltre, sulla base dei risultati ottenuti è possibile generare diversi nodi di preparazione dati. Per esempio, è possibile generare un nodo Filtro che escluda i campi che presentano troppi valori mancanti per essere utili ai fini della modellazione e generare un Supernodo che assegni i valori mancanti ad alcuni o tutti i campi restanti. In questi casi viene evidenziata la reale potenza dell'esplorazione, che consente non solo di valutare lo stato corrente dei propri dati, ma anche di eseguire azioni in base alla valutazione.

Screening o campionamento dei dati. Poiché un'esplorazione iniziale è particolarmente efficace quando vengono utilizzate "grandi quantità di dati," è possibile utilizzare un nodo Campione per ridurre il tempo di elaborazione durante l'esplorazione iniziale selezionando solo un sottoinsieme di record. Il nodo Esplora può anche essere utilizzato assieme ad altri nodi, per esempio i nodi Selezione funzioni e Rilevamento anomalie, durante le fasi preliminari dell'analisi.

Scheda Impostazioni del nodo Esplora

La scheda Impostazioni consente di specificare i parametri di base per la verifica.

Default. È possibile collegare semplicemente il nodo al flusso e fare clic su **Esegui** per generare un report di verifica per tutti i campi basato sulle impostazioni di default, nel modo seguente:

- In assenza di impostazioni del nodo Tipo, nel report vengono inclusi tutti i campi.
- In presenza di impostazioni del nodo Tipo, indipendentemente dalla relativa istanziazione, nella visualizzazione vengono inclusi tutti i campi *Input*, *Obiettivo* ed *Entrambi*. Se è disponibile un solo campo *Obiettivo*, questo viene utilizzato come campo di sovrapposizione. Se sono disponibili più campi *Obiettivo*, non è specificata una sovrapposizione di default.

Utilizza campi di default. Selezionare questa opzione per selezionare i campi manualmente. Utilizzare il pulsante selettore di campo riportato a destra per selezionare i campi singolarmente o per tipo.

Campo sovrapposto. Il campo sovrapposto viene utilizzato per disegnare i grafici in miniatura visualizzati nel report di verifica. Nel caso di un campo continuo (intervallo numerico), vengono calcolate anche le statistiche bivariate (covarianza e correlazione). Se è presente un solo campo *Obiettivo* in base alle impostazioni del nodo Tipo, tale campo viene utilizzato come campo di sovrapposizione di default, come descritto in precedenza. In alternativa, è possibile selezionare **Utilizza campi di default** per specificare una sovrapposizione.

Visualizzazione. Consente di specificare se i grafici sono disponibili nell'output e di scegliere le statistiche visualizzate per default.

- **Grafici.** Visualizza un grafico per ogni campo selezionato; a seconda dei dati, tale grafico può essere di distribuzione (a barre), un istogramma o un grafico a dispersione. I grafici vengono visualizzati come miniature nel report iniziale, ma è anche possibile generare grafici a schermo intero e nodi di grafici. Per ulteriori informazioni, consultare l'argomento "Browser di output del nodo Esplora" a pagina 285.

- **Statistiche di base/Statistiche avanzate.** Specifica il livello delle statistiche visualizzate per default nell'output. Questa impostazione determina la visualizzazione iniziale, ma tutte le statistiche sono disponibili nell'output, indipendentemente da questa impostazione. Per ulteriori informazioni, consultare l'argomento "Visualizza statistiche" a pagina 286.

Calcola mediana e moda. Calcola la mediana e la moda di tutti i campi nel report. Nel caso di insiemi di dati di grandi dimensioni, queste statistiche potrebbero rallentare l'elaborazione in quanto richiedono più tempo di altre per il calcolo. Per quanto riguarda la mediana, il valore riportato, in alcuni casi, potrebbe essere basato su un campione di 2000 record anziché su tutto l'insieme di dati. Questo campionamento viene eseguito per ogni campo nei casi in cui si rischierebbe di superare i limiti di memoria. Quando è attivo il campionamento, i risultati verranno etichettati di conseguenza nell'output (*Mediana campione* invece che semplicemente *Mediana*). Tutte le altre statistiche vengono invece sempre calcolate utilizzando l'intero insieme di dati.

Campi vuoti o senza tipo. Quando sono utilizzati con i dati istanziati, i campi senza tipo non sono inclusi nel report di verifica. Per includere i campi senza tipo (compresi quelli vuoti), selezionare **Cancella tutti i valori** nei nodi Tipo a monte. Questa operazione fa in modo che i dati non vengano istanziati, determinando l'inclusione di tutti i campi nel report. Questa funzione può essere utile, per esempio, se si desidera ottenere un elenco completo di tutti i campi o generare un nodo Filtro che escluda i campi vuoti. Per ulteriori informazioni, consultare l'argomento "Filtraggio dei campi con dati mancanti" a pagina 289.

Nodo Esplora, scheda Qualità

La scheda Qualità del nodo Esplora fornisce le opzioni per la gestione di valori mancanti, anomali ed estremi.

Valori mancanti

- **Conteggio dei record con valori validi.** Selezionare questa opzione per mostrare il numero di record con valori validi per ogni campo valutato. Si noti che i valori null (non definiti), i valori vuoti, gli spazi e le stringhe vuote vengono sempre trattati come valori non validi.
- **Conteggi distinti dei record con valori non validi.** Selezionare questa opzione per mostrare il numero di record con ogni tipo di valore non valido per ciascun campo.

Valori anomali ed estremi

Metodo di rilevamento dei valori anomali ed estremi. Sono supportati due metodi:

Deviazione standard dalla media. Rileva i valori anomali ed estremi in base al numero di deviazioni standard dalla media. Per esempio, se un campo ha una media di 100 e una deviazione standard di 10, è possibile specificare 3.0 per indicare che tutti i valori inferiori a 70 o superiori a 130 devono essere considerati anomali.

Intervallo interquartile. Rileva i valori anomali ed estremi in base all'intervallo interquartile, vale a dire l'intervallo in cui rientrano i due quartili centrali (fra il 25mo e il 75mo percentile). Ad esempio, in base all'impostazione di default 1.5, la soglia inferiore per i valori anomali è $Q1 - 1.5 * IQR$ e la soglia superiore è $Q3 + 1.5 * IQR$. Si noti che l'utilizzo di questa opzione su insiemi di dati di grandi dimensioni può rallentare la performance.

Browser di output del nodo Esplora

Il browser del nodo Esplora è uno strumento particolarmente efficace per ottenere una panoramica dei propri dati. La scheda Esplora visualizza grafici in miniatura, icone di archiviazione e statistiche per tutti i campi, mentre la scheda Qualità visualizza informazioni sui valori anomali, estremi e mancanti. In base ai grafici iniziali e alle statistiche riassuntive, è possibile decidere di ricodificare un campo numerico o i valori di un campo nominale oppure di derivare un nuovo campo. In alternativa, può rivelarsi utile

approfondire l'esplorazione utilizzando una visualizzazione più avanzata. A questo scopo, dal browser del report di verifica, utilizzare il menu Genera per creare un numero qualsiasi di nodi che possono essere utilizzati per trasformare o visualizzare i dati.

- Ordinare le colonne facendo clic sulla relativa intestazione oppure trascinandole e rilasciandole. È supportata anche la maggior parte delle operazioni di output. Per ulteriori informazioni, consultare l'argomento "Visualizzazione dell'output" a pagina 272.
- Visualizzare valori e intervalli per i campi facendo doppio clic sulle colonne Misurazione o Univocità di un campo.
- Utilizzare la barra degli strumenti o il menu Modifica per visualizzare o nascondere le etichette dei valori oppure per scegliere le statistiche da visualizzare. Per ulteriori informazioni, consultare l'argomento "Visualizza statistiche".
- Verificare le icone di archiviazione a sinistra dei nomi dei campi. L'archiviazione definisce le modalità di archiviazione dei dati in un campo. Per esempio, un campo con valori 1 e 0 è un campo in cui sono archiviati dati di tipo intero. Si distingue dal livello di misurazione, che descrive l'utilizzo dei dati e non influisce sull'archiviazione. Per ulteriori informazioni, consultare l'argomento "Impostazione dell'archiviazione e della formattazione dei campi" a pagina 24.

Visualizzazione e generazione dei grafici

Se non è selezionata una sovrapposizione, la scheda Esplora visualizza i grafici a barre (per campi nominali o flag) o gli istogrammi (campi continui).

In caso di sovrapposizione di un campo nominale o flag, i grafici vengono colorati in base ai valori della sovrapposizione.

In caso di sovrapposizione di un campo continuo, anziché barre e istogrammi monodimensionali vengono generati grafici a dispersione bidimensionali. In questo caso, l'asse x viene mappato al campo di sovrapposizione, consentendo di visualizzare la stessa scala su tutti gli assi x mentre si procede con la lettura della tabella.

- Per i campi flag o nominali, tenere il cursore del mouse su una barra in modo da visualizzare l'etichetta o il valore sottostante in una descrizione.
- Per i campi flag o nominali, utilizzare la barra degli strumenti per cambiare da orizzontale a verticale l'orientamento dei grafici in miniatura.
- Per generare un grafico a schermo intero da una miniatura, fare doppio clic su di essa oppure selezionare una miniatura e scegliere **Output grafico** dal menu Genera. *Nota:* se era stato creato un grafico in miniatura in base a dati campionati e il flusso di dati originale è ancora aperto, il grafico generato conterrà tutti i casi.

Il grafico si può generare solo se il nodo Esplora che ha generato l'output è collegato al flusso.

- Per generare un nodo di grafico corrispondente, selezionare uno o più campi nella scheda Esplora e scegliere **Nodo Grafico** nel menu Genera. Il nodo risultante viene aggiunto all'area del flusso e può essere utilizzato per ricreare il grafico ogni volta che viene eseguito il flusso.
- se un insieme di sovrapposizione contiene più di 100 valori, verrà visualizzato un avviso e non verrà inclusa la sovrapposizione.

Visualizza statistiche

La finestra di dialogo Visualizza statistiche consente di scegliere le statistiche visualizzate nella scheda Esplora. Le impostazioni iniziali vengono specificate nel nodo Esplora. Per ulteriori informazioni, consultare l'argomento "Scheda Impostazioni del nodo Esplora" a pagina 284.

Minimo. Il valore più basso di una variabile numerica.

Massimo. Il valore più alto di una variabile numerica.

Somma. La somma o il totale di tutti i valori non mancanti di tutti i casi.

Intervallo. La differenza tra il valore massimo ed il valore minimo di una variabile numerica.

Media. Una misura di tendenza centrale. Media aritmetica, ovvero somma divisa per il numero di casi.

Errore standard della media. Una misura di quanto il valore della media può variare tra campioni presi dalla stessa distribuzione. Può essere utilizzata per confrontare genericamente la media osservata rispetto a un valore ipotizzato (ovvero, è possibile concludere che i due valori sono diversi se il rapporto della differenza rispetto all'errore standard è inferiore a -2 o maggiore di +2).

Deviazione standard. Una misura di dispersione intorno alla media, uguale alla radice quadrata della varianza. L'unità di misura della deviazione standard è la stessa della variabile originale.

Varianza. Una misura della dispersione dei valori intorno alla media. È calcolata come somma dei quadrati degli scostamenti dalla media, divisa per un valore corrispondente alla somma dei casi meno uno. La varianza è espressa in quadrati dell'unità di misura della variabile.

Asimmetria. Una misura dell'asimmetria di una distribuzione. La distribuzione normale è simmetrica e ha un valore di asimmetria pari a 0. Una distribuzione con una notevole asimmetria positiva ha una lunga coda a destra. Una distribuzione con asimmetria negativa ha una coda a sinistra. In generale un'asimmetria con valore più che doppio dell'errore standard indica lo scostamento dalla normale simmetria.

Errore standard dell'asimmetria. Il rapporto fra la simmetria di una distribuzione e il suo errore standard viene usato come test di normalità. L'ipotesi di normalità può essere rifiutata se questo rapporto è inferiore di 2 o maggiore di +2. Un valore positivo elevato per l'asimmetria indica una coda a destra lunga; un valore negativo estremo indica una coda a sinistra lunga.

Curtosi. Una misura di quanto le osservazioni si trovino raggruppate in un punto centrale. Per la distribuzione normale, il valore della statistica di curtosi è zero. Una curtosi positiva indica che, rispetto a una distribuzione normale, le osservazioni sono più raggruppate intorno al centro della distribuzione e hanno code più sottili fino ai valori estremi della distribuzione; a quel punto, le code della distribuzione leptocurtica sono più spesse rispetto a una distribuzione normale. Una curtosi negativa indica che, rispetto a una distribuzione normale, le osservazioni sono meno raggruppate e hanno code più spesse fino ai valori estremi della distribuzione; a quel punto, le code della distribuzione platicurtica sono più sottili rispetto a una distribuzione normale.

Errore standard della curtosi. Il rapporto fra la curtosi di una distribuzione e il suo errore standard viene usato come test di normalità. L'ipotesi di normalità può essere rifiutata se questo rapporto è inferiore di 2 o maggiore di +2. Un valore positivo elevato per la curtosi indica che le code della distribuzione sono più lunghe di quelle di una distribuzione normale; un valore negativo per la curtosi indica code più corte, simili a quelle di una distribuzione uniforme a forma di scatola.

Univocità. Valuta tutti gli effetti simultaneamente, correggendo ogni effetto per tutti gli altri effetti di qualunque tipo.

Validi. I casi validi il cui valore non è né il valore mancante di sistema né un valore definito come mancante dall'utente.

Median. È il valore sopra il quale e sotto il quale ricade la metà dei casi, il 50-esimo percentile. Se il numero di casi è pari, la mediana è pari alla media dei due casi centrali quando questi sono ordinati secondo l'ordine ascendente o discendente. La mediana è una misura di tendenza centrale non sensibile ai valori anomali, a differenza della media che può essere influenzata da valori eccezionalmente bassi o alti.

Modalità. Il valore che ricorre più frequentemente. Se più valori condividono la maggiore ricorrenza, ognuno di essi è una modalità.

Si noti che mediana e moda sono escluse per default al fine di migliorare la performance; tuttavia, è possibile selezionarle nella scheda Impostazioni del nodo Esplora. Per ulteriori informazioni, consultare l'argomento "Scheda Impostazioni del nodo Esplora" a pagina 284.

Statistiche per le sovrapposizioni

Se viene utilizzato un campo di sovrapposizione continuo (intervallo numerico), sono disponibili anche le seguenti statistiche:

Covarianza. Una misura non standardizzata di associazione tra due variabili, pari alla deviazione del prodotto incrociato divisa per N-1.

Browser Esplora, scheda Qualità

La scheda Qualità nel browser del nodo Esplora visualizza i risultati dell'analisi della qualità dei dati e consente di specificare i trattamenti per valori anomali, estremi e valori mancanti.

Assegnazione dei valori mancanti: Il report di verifica elenca la percentuale dei record completi per ciascun campo, nonché il numero dei valori validi, null e vuoti. È possibile assegnare i valori mancanti a campi specifici secondo le proprie esigenze e quindi generare un Supernodo per applicare queste trasformazioni.

1. Nella colonna **Assegna valori mancanti**, specificare il tipo di valori che si desidera assegnare. Si può decidere di assegnare i valori vuoti, null o entrambi, oppure è possibile specificare una condizione o un'espressione personalizzata per selezionare i valori da assegnare.

Esistono vari tipi di valori mancanti riconosciuti da IBM SPSS Modeler:

- **Valori null o mancanti di sistema.** Sono valori non stringa che vengono lasciati vuoti nel database o nel file di origine e che non sono specificati in modo esplicito come "mancanti" in un nodo Tipo o in un nodo origine. I valori mancanti di sistema vengono visualizzati come **\$null\$**. In IBM SPSS Modeler le stringhe vuote non vengono considerate valori null, ma possono essere considerate valori null da alcuni database.
- **Stringhe vuote e spazi vuoti.** I valori stringa vuote e spazi vuoti (stringhe prive di caratteri visibili) sono trattati in modo diverso rispetto ai valori null. Nella maggior parte dei casi, le stringhe vuote vengono considerate equivalenti agli spazi vuoti. Per esempio, se si seleziona l'opzione che consente di trattare gli spazi vuoti come valori vuoti in un nodo origine Tipo, questa impostazione risulta valida anche per le stringhe vuote.
- **Valori vuoti o mancanti definiti dall'utente.** Sono valori come, ad esempio, sconosciuto, 99 o -1 definiti in modo esplicito come mancanti in un nodo origine o un nodo Tipo. Se lo si desidera, si può anche decidere di considerare i valori null e gli spazi vuoti come valori vuoti: in questo modo è possibile evidenziarli in modo da poter essere elaborati diversamente ed esclusi dalla maggior parte dei calcoli. Per esempio, è possibile utilizzare la funzione @BLANK per trattare questi valori, insieme ad altri tipi di valori mancanti, come valori vuoti.

2. Nella colonna **Metodo**, indicare il metodo da utilizzare.

Per l'immissione di valori mancanti sono disponibili i seguenti metodi:

Fisso. Sostituisce con un valore fisso (la media di un campo, il valore intermedio dell'intervallo o una costante specificata dall'utente).

Casuale. Sostituisce con un valore casuale in base a una distribuzione normale o uniforme.

Espressione. Consente di specificare un'espressione personalizzata. Per esempio, è possibile sostituire valori con una variabile globale creata dal nodo Calcola globali.

Algoritmo. Sostituisce un valore previsto da un modello basato sull'algoritmo C&RT. Per ciascun campo immesso utilizzando questo metodo, sarà presente un modello C&RT separato, insieme ad un nodo Riempimento che sostituisce i valori vuoti e null con il valore previsto dal modello. Viene quindi utilizzato un nodo Filtro per rimuovere i campi di previsione generati dal modello.

3. Per generare un Supernodo valori mancanti, dai menu scegliere:

Genera > Supernodo valori mancanti

Viene visualizzata la finestra di dialogo Supernodo valori mancanti.

4. Selezionare **Tutti i campi** o **Solo campi selezionati** e, se lo si desidera, specificare la dimensione del campione (il campione specificato è un valore percentuale; per default viene campionato il 10% di tutti i record).
5. Fare clic su **OK** per aggiungere il Supernodo generato all'area del flusso.
6. Collegare il Supernodo al flusso in modo da applicare le trasformazioni.

Nel Supernodo viene utilizzata, a seconda dei casi, una combinazione di nugget del modello, nodi Riempimento e Filtro. Per comprenderne il funzionamento, è possibile modificare il Supernodo e fare clic su **Zoom avanti**; inoltre, è possibile aggiungere, modificare o rimuovere nodi specifici all'interno del Supernodo per modificarne il comportamento secondo le esigenze.

Gestione dei valori anomali ed estremi: Il report di verifica elenca il numero dei valori anomali ed estremi per ciascun campo in base alle opzioni di rilevamento specificate nel nodo Verifica dati. Per ulteriori informazioni, consultare l'argomento "Nodo Esplora, scheda Qualità" a pagina 285. È possibile decidere di forzare, scartare o annullare questi valori per campi specifici in funzione delle proprie esigenze, e quindi generare un Supernodo per applicare le trasformazioni.

1. Nella colonna **Azione**, indicare in che modo si desidera gestire i valori anomali ed estremi per campi specifici.

Per la gestione di valori anomali ed estremi sono disponibili le seguenti azioni:

- **Forza.** Consente di sostituire i valori anomali ed estremi con il valore più vicino che non sia considerato estremo. Per esempio, se un valore viene definito anomalo quando è superiore o inferiore a 3 deviazioni standard, tutti i valori anomali vengono sostituiti con il valore più alto o più basso di questo intervallo.
- **Scarta.** Consente di scartare i record con valori anomali o estremi per il campo specificato.
- **Annulla.** Sostituisce i valori anomali ed estremi con il valore null o mancante di sistema.
- **Forza valori anomali / scarta estremi.** Scarta soltanto i valori estremi.
- **Forza valori anomali / annulla estremi.** Annulla soltanto i valori estremi.

2. Per generare il Supernodo, dai menu scegliere:

Genera > Supernodo valori anomali ed estremi

Viene visualizzata la finestra di dialogo Supernodo Valori anomali.

3. Selezionare **Tutti i campi** o **Solo campi selezionati**, quindi fare clic su **OK** per aggiungere il Supernodo generato all'area del flusso.
4. Collegare il Supernodo al flusso in modo da applicare le trasformazioni.

Se lo si desidera, è possibile modificare il Supernodo e ingrandirlo per esaminarlo o modificarlo. All'interno del Supernodo, i valori vengono scartati, forzati o annullati mediante una serie di nodi Seleziona e/o Riempimento, secondo necessità.

Filtraggio dei campi con dati mancanti: Dal browser del nodo Esplora, è possibile creare un nuovo Filtro in base ai risultati dell'analisi Qualità utilizzando la finestra di dialogo Genera filtro da qualità.

Modalità. Selezionare l'operazione desiderata per i campi specificati, ovvero **Includi** o **Escludi**.

- **Campi selezionati.** Il nodo Filtro includerà/escluderà i campi selezionati nella scheda Qualità. Per esempio, è possibile ordinare la tabella in base alla colonna % **Complete**, utilizzare il metodo Maiusc-clic per selezionare i campi con la percentuale di completamento più bassa e generare un nodo Filtro che escluda tali campi.
- **Campi con percentuale di qualità maggiore di.** Il nodo Filtro includerà/escluderà i campi in cui la percentuale di record completi è superiore alla soglia specificata. La soglia di default è 50%.

Filtraggio dei campi vuoti o senza tipo

Dopo che sono stati istanziati i valori dei dati, i campi senza tipo o vuoti vengono esclusi dai risultati dell'esplorazione e dalla maggior parte degli output in IBM SPSS Modeler. Questi campi vengono ignorati ai fini della modellazione, ma potrebbero "sovraffollare" e confondere inutilmente i dati. In questi casi, è possibile utilizzare il browser del nodo Esplora per generare un nodo Filtro che rimuova questi campi dal flusso.

1. Per assicurarsi che tutti i campi vengano inclusi nell'esplorazione, compresi i campi vuoti o senza tipo, selezionare **Cancella tutti i valori** nel nodo origine o nel nodo Tipo a monte o impostare i Valori su *<Passa>* per tutti i campi.
2. Nel browser del nodo Esplora, ordinare in base alla colonna **% Complete**, selezionare i campi che hanno zero valori validi (o qualche altra soglia) e utilizzare il menu Genera per creare un nodo Filtro che possa essere aggiunto al flusso.

Selezione di record con dati mancanti: Dal browser del nodo Esplora, è possibile creare un nuovo nodo Seleziona in base ai risultati dell'analisi Qualità.

1. Dal browser del nodo Esplora, selezionare la scheda Qualità.
2. Dal menu, scegliere:

Genera > Valori mancanti nodo Seleziona

Viene visualizzata la finestra di dialogo Genera nodo Selezione.

Seleziona se il record è: Specificare se conservare i record quando sono di tipo **Valido** o **Non valido**.

Cerca valori non validi in. Specificare se cercare valori non validi.

- **Tutti i campi.** Il nodo Seleziona controlla tutti i campi per verificare se contengono valori non validi.
- **Campi selezionati nella tabella.** Il nodo Seleziona esegue il controllo solo nei campi selezionati nella tabella di output Qualità.
- **Campi con percentuale di qualità maggiore di.** Il nodo Seleziona esegue il controllo nei campi in cui la percentuale di record completi è maggiore della soglia specificata. La soglia di default è 50%.

Considera un record non valido se è presente un valore non valido in. Specificare la condizione per identificare un record come non valido.

- **Qualsiasi campo precedente.** Il nodo Seleziona considererà un record non valido se uno *qualsiasi* dei campi specificati sopra contiene un valore non valido per tale record.
- **Tutti i campi precedenti.** Il nodo Seleziona considererà un record non valido solo se *tutti* i campi specificati sopra contengono valori non validi per tale record.

Generazione di altri nodi per la preparazione dei dati

Numerosi nodi utilizzabili nella fase di preparazione dei dati (quali i nodi Ricodifica, Raccolta e Ricava) possono essere generati direttamente dal browser del nodo Esplora. Ad esempio:

- È possibile derivare un nuovo campo in base ai valori di *valore_richiesta* e *reddito*, selezionando entrambi nel report di verifica e scegliendo **Ricava** dal menu Genera. Il nuovo nodo verrà aggiunto all'area del flusso.
- Analogamente, in base ai risultati dell'esplorazione si potrebbe determinare che la ricodifica di *reddito* in bin basati su percentili fornisce un'analisi più mirata. Per generare un nodo Discretizza, selezionare la riga del campo nella visualizzazione e scegliere **Discretizza** dal menu Genera.

Dopo la generazione del nodo e la relativa aggiunta all'area del flusso, è necessario collegare il nodo al flusso e aprirlo per specificare le opzioni per i campi selezionati.

Nodo Trasformazioni

La normalizzazione dei campi di input rappresenta un'operazione importante prima dell'utilizzo delle tecniche di calcolo dei punteggi tradizionali, quali regressione, regressione logistica e analisi discriminante. Queste tecniche implicano delle congetture relative alle normali distribuzioni dei dati che potrebbero non essere vere per molti file di dati grezzi. Un modo per gestire i dati reali consiste nell'applicare trasformazioni che avvicinano un elemento di dati grezzo a una distribuzione più normale. Inoltre, i campi normalizzati possono essere facilmente confrontati tra loro — ad esempio, reddito ed età si trovano su scale completamente differenti in un file di dati non elaborato ma, quando viene eseguita la normalizzazione, l'impatto relativo di ciascun campo può essere facilmente interpretato.

Il nodo Trasformazioni è dotato di un Visualizzatore output che consente di valutare visivamente e rapidamente la migliore trasformazione da utilizzare. A prima vista è possibile capire se le variabili sono distribuite normalmente e, quando necessario, scegliere la trasformazione desiderata e applicarla. È possibile scegliere più campi ed eseguire una trasformazione per campo.

Dopo avere selezionato le trasformazioni desiderate per i campi, è possibile generare nodi Ricava o Riempimento che eseguono le trasformazioni e collegare questi nodi al flusso. Il nodo Ricava crea campi nuovi, mentre il nodo Riempimento trasforma i campi esistenti. Per ulteriori informazioni, consultare l'argomento "Generazione di grafici" a pagina 293.

Scheda Campi del nodo Trasformazioni

Nella scheda Campi è possibile specificare quali campi dei dati utilizzare per visualizzare le possibili trasformazioni e applicarle. Solo i campi numerici possono essere trasformati. Fare clic sul pulsante del selettore di campo e selezionare uno o più campi numerici dall'elenco visualizzato.

Scheda Opzioni del nodo Trasformazioni

La scheda Opzioni consente di specificare il tipo di trasformazioni che si desidera includere. Si può decidere di includere tutte le trasformazioni disponibili, oppure selezionare le singole trasformazioni.

In questo secondo caso, è anche possibile immettere un numero per distanziare i dati per le trasformazioni inversa e logaritmica. Questa operazione è utile nei casi in cui un'alta proporzione di zeri nei dati potrebbe condizionare i risultati di media e deviazione standard.

Per esempio, si supponga di avere un campo denominato *SALDO* che contiene alcuni valori zero e di volervi applicare la trasformazione inversa. Per evitare alterazioni e condizionamenti indesiderati, in questo caso sarebbe opportuno selezionare **Inverso (1/x)** e immettere 1 nel campo **Utilizza un offset tra i dati**. (Questo offset non è correlato a quello applicato dalla funzione di sequenza @OFFSET in IBM SPSS Modeler.)

Tutte le formule. Indica che tutte le trasformazioni disponibili devono essere calcolate e visualizzate nell'output.

Seleziona formule. Consente di selezionare le diverse trasformazioni da calcolare e mostrare nell'output.

- **Inverso (1/x).** Indica che la trasformazione inversa deve essere visualizzata nell'output.
- **Logaritmo (log n).** Indica che nell'output deve essere visualizzata la trasformazione logaritmica n .
- **Logaritmo (log 10).** Indica che la trasformazione logaritmica $_{10}$ deve essere visualizzata nell'output.
- **Esponenziale.** Indica che la trasformazione esponenziale (e^x) deve essere visualizzata nell'output.
- **Radice quadrata.** Indica che la trasformazione radice quadrata deve essere visualizzata nell'output.

Scheda Output del nodo Trasformazioni

La scheda Output consente di specificare il formato e la posizione dell'output. Si può scegliere di visualizzare i risultati sullo schermo o di inviarli a uno dei tipi di file standard. Per ulteriori informazioni, consultare l'argomento "Scheda Output del nodo Output" a pagina 276.

Visualizzatore output del nodo Trasformazioni

Il Visualizzatore output permette di visualizzare i risultati dell'esecuzione del nodo Trasformazioni. Si tratta di uno strumento molto efficace che visualizza più trasformazioni per ogni campo sotto forma di miniature, consentendo un rapido confronto visivo dei campi. Per salvare, esportare o stampare l'output, utilizzare le relative opzioni nel menu File. Per ulteriori informazioni, consultare l'argomento "Visualizzazione dell'output" a pagina 272.

Sotto ogni trasformazione, ad eccezione di Trasformazione selezionata, viene visualizzata una legenda nel formato:

Media (Deviazione standard)

Generazione di nodi per le trasformazioni

Il Visualizzatore output costituisce un'utile base di partenza per la fase di preparazione dei dati. Per esempio, si può decidere di normalizzare il campo *ETÀ* in modo da poter utilizzare una tecnica di calcolo del punteggio (come la regressione logistica o l'analisi discriminante) che presuppone una distribuzione normale. In base ai grafici iniziali e alle statistiche riassuntive, è possibile decidere di trasformare il campo *ETÀ* secondo una distribuzione specifica (per esempio, logaritmica). Dopo avere selezionato la distribuzione desiderata, è possibile generare un nodo Ricava con una trasformazione standardizzata da utilizzare per il calcolo del punteggio.

Dal Visualizzatore output è possibile creare i seguenti nodi Operazioni su campi:

- Ricava
- Riempimento

Un nodo Ricava crea campi nuovi con le trasformazioni desiderate, mentre il nodo Riempimento trasforma i campi esistenti. I nodi vengono posti nell'area sotto forma di Supernodo.

Se si seleziona la stessa trasformazione per diversi campi, un nodo Ricava o Riempimento contiene le formule per quel tipo di trasformazione per tutti i campi a cui si applica quella trasformazione. Ad esempio, si supponga che siano stati selezionati i campi e le trasformazioni, visualizzati nella seguente tabella, per generare un nodo Ricava.

Tabella 50. Esempio di generazione di nodo Ricava.

Campo	Trasformazione
<i>AGE</i>	Distribuzione corrente
<i>INCOME</i>	Registro
<i>OPEN_BAL</i>	Inverso
<i>BALANCE</i>	Inverso

Il Supernodo contiene i nodi seguenti:

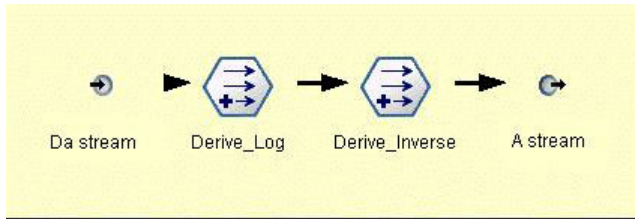


Figura 66. Supernodo nell'area

In questo esempio, il nodo Ricava_Logaritmica presenta la formula logaritmica per il campo *REDDITO*, mentre il nodo Ricava_Inversa presenta le formule inverse per i campi *SALDO_APERT* e *SALDO*.

Per generare un nodo

1. Selezionare la trasformazione desiderata per ciascun campo nel Visualizzatore output.
2. Dal menu Genera, scegliere **Nodo Ricava** o **Nodo Riempimento**, secondo necessità.

Viene visualizzata la finestra di dialogo Genera nodo Ricava o Genera nodo Riempimento, a seconda di quanto selezionato.

Scegliere **Trasformazione non standardizzata** o **Trasformazione standardizzata (punteggio z)**, secondo necessità. La seconda opzione applica un punteggio z alla trasformazione; i punteggi z rappresentano i valori come funzione della distanza della variabile dalla media nelle deviazioni standard. Per esempio, se si applica la trasformazione logaritmica al campo *ETÀ* e si sceglie la trasformazione standardizzata, l'equazione finale del nodo generato sarà:

$$(\log(\text{AGE}) - \text{Mean}) / \text{SD}$$

Dopo che il nodo è stato generato ed è visualizzato nell'area del flusso:

1. Collegarlo al flusso.
2. Nel caso di un Supernodo, se lo si desidera, è possibile fare doppio clic per visualizzarne il contenuto.
3. Se necessario, fare doppio clic sul nodo Ricava o Riempimento per modificare le opzioni relative ai campi selezionati.

Generazione di grafici: È possibile generare output di istogrammi a schermo intero da un istogramma in miniatura nel Visualizzatore output.

Per generare un grafico

1. Fare doppio clic su un grafico in miniatura nel Visualizzatore output.

oppure

Selezionare un grafico in miniatura nel Visualizzatore output.

2. Dal menu Genera, scegliere **Output grafico**.

Viene visualizzato l'istogramma con una normale curva di distribuzione sovrapposta. Questo consente di confrontare l'effettiva corrispondenza tra ogni trasformazione disponibile e una normale distribuzione.

Nota: è possibile generare un grafico solo se il nodo Trasformazioni che ha creato l'output è connesso al flusso.

Altre operazioni: Dal Visualizzatore output è anche possibile:

- Ordinare la griglia di output in base alla colonna Campo.
- Esportare l'output in un file HTML. Per ulteriori informazioni, consultare l'argomento "Esportazione dell'output" a pagina 274.

Nodo Statistiche

Il nodo Statistiche fornisce informazioni riassuntive di base su campi numerici. È possibile ottenere statistiche riassuntive per singoli campi e per correlazioni tra campi.

Scheda Impostazioni del nodo Statistiche

Esamina. Selezionare il campo o i campi per i quali si desiderano statistiche riassuntive individuali. È possibile selezionare più campi.

Statistiche. Selezionare le statistiche da inserire nel report. Le opzioni disponibili includono **Conteggio**, **Media**, **Somma**, **Min**, **Max**, **Intervallo**, **Varianza**, **Dev. std**, **Errore std di media**, **Mediana** e **Modalità**.

Correla. Selezionare il campo o i campi che si desidera correlare. È possibile selezionare più campi. Una volta selezionati i campi di correlazione, nell'output verrà elencata la correlazione tra ogni campo Esamina e il campo o i campi di correlazione.

Impostazioni di correlazione. È possibile impostare una serie di opzioni per visualizzare l'intensità delle correlazioni nell'output.

Impostazioni di correlazione

IBM SPSS Modeler è in grado di caratterizzare le correlazioni tramite etichette descrittive in modo da sottolineare le relazioni importanti. La **correlazione** misura l'intensità della relazione tra due campi continui (intervallo numerico). Accetta valori compresi tra -1.0 e 1.0 . I valori vicini a $+1.0$ indicano una forte associazione positiva: in questo caso, i valori alti in un campo sono associati ai valori alti nell'altro campo e i valori bassi sono associati ai valori bassi. I valori vicini a -1.0 indicano una forte associazione negativa in modo che i valori alti per un campo vengono associati ai valori bassi per l'altro campo e viceversa. Valori vicini a 0.0 indicano un'associazione debole, in questo caso i valori nei due campi sono più o meno indipendenti.

Utilizzando la finestra di dialogo Impostazioni di correlazione, è possibile controllare la visualizzazione delle etichette di correlazione, modificare le soglie che definiscono le categorie e modificare le etichette utilizzate per ciascun intervallo. Dal momento che la caratterizzazione dei valori di correlazione dipende in modo rilevante dall'ambito del problema, è possibile che si desideri personalizzare gli intervalli e le etichette in base a una specifica situazione.

Visualizza etichette dell'intensità di correlazione nell'output. Questa opzione è selezionata per default. Deselezionarla per omettere le etichette descrittive nell'output.

Intensità di correlazione. Esistono due opzioni per definire e identificare l'intensità delle correlazioni:

- **Definisci l'intensità di correlazione in base all'importanza (1-p).** Le correlazioni vengono identificate in base all'importanza, definita come 1 meno la significatività o 1 meno la probabilità che la differenza di media possa essere spiegata solo dal caso. Quanto più questo valore si avvicina a 1 , maggiore è la possibilità che i due campi *non* siano indipendenti — in altre parole, che esista una relazione tra i due campi. L'assegnazione di etichette alle correlazioni in base all'importanza è generalmente consigliata rispetto al valore assoluto perché considera la variabilità dei dati — ad esempio, un coefficiente 0.6 può essere estremamente significativo in un insieme di dati e non significativo in un altro insieme di dati. Per default, un'importanza con valore compreso tra 0.0 e 0.9 viene definita di tipo *Debole*, una con valore compreso tra 0.9 e 0.95 viene definita di tipo *Medio* e una con valore compreso tra 0.95 e 1.0 viene definita di tipo *Forte*.
- **Definisci l'intensità di correlazione in base al valore assoluto.** Le correlazioni sono basate sul valore assoluto del coefficiente di correlazione di Pearson, che è compreso tra -1 e 1 , come descritto in precedenza. Più il valore assoluto di questa misura si avvicina a 1 , più intensa è la correlazione. Per default, le correlazioni tra 0.0 e 0.3333 (in valore assoluto) sono etichettate come di tipo *Debole*, quelle tra 0.3333 e 0.6666 sono etichettate come di tipo *Medio* e quelle tra 0.6666 e 1.0 sono etichettate come di tipo *Forte*. Si noti, tuttavia, che è difficile generalizzare la significatività di qualsiasi valore da un

insieme di dati a un altro; per questo motivo, nella maggior parte dei casi è consigliabile definire le correlazioni in base alla probabilità anziché in base al valore assoluto.

Browser di output del nodo Statistiche

Il browser di output del nodo Statistiche visualizza i risultati dell'analisi statistica e consente di eseguire operazioni tra cui la selezione dei campi, la generazione di nuovi nodi in base alla selezione ed il salvataggio e la stampa dei risultati. Le normali opzioni di salvataggio, esportazione e stampa sono disponibili dal menu File e le normali opzioni di modifica sono disponibili dal menu Modifica. Per ulteriori informazioni, consultare l'argomento "Visualizzazione dell'output" a pagina 272.

Quando si visualizza per la prima volta l'output del nodo Statistiche, i risultati vengono espansi. Per nascondere i risultati, utilizzare il controllo dell'espansore di un elemento se si desidera comprimere solo i risultati di tale elemento oppure fare clic sul pulsante **Comprimi tutto** se si desidera nascondere tutti i risultati. Per visualizzare nuovamente i risultati dopo averli compressi, utilizzare il controllo dell'espansore a sinistra di un elemento per visualizzarne i risultati, oppure fare clic sul pulsante **Espandi tutto** per visualizzare tutti i risultati.

L'output include una sezione per ogni campo *Esamina*, contenente una tabella delle statistiche richieste.

- **Conteggio.** Numero di record con valori validi per il campo.
- **Media.** Valore medio (media) del campo in tutti i record.
- **Somma.** Somma dei valori del campo in tutti i record.
- **Min.** Valore minimo del campo.
- **Max.** Valore massimo del campo.
- **Intervallo.** Differenza tra i valori minimi e massimi.
- **Varianza.** Misura della variabilità dei valori di un campo. Viene ottenuta calcolando la differenza tra ogni valore e la media globale, elevandola al quadrato, sommando tutti i valori e dividendo il risultato per il numero di record.
- **Deviazione standard.** Ulteriore misura della variabilità nei valori di un campo, calcolata come radice quadrata della varianza.
- **Errore standard della media.** Misura dell'incertezza nella stima della media del campo. Tale misura risulta utile se si presume che la media venga applicata a dati nuovi.
- **Mediana.** Valore "centrale" del campo, ovvero il valore che divide la metà superiore dei dati dalla metà inferiore (in base ai valori del campo).
- **Modalità.** Valore singolo più comune nei dati.

Correlazioni. Se sono stati specificati campi correlati, l'output contiene inoltre una sezione che elenca la correlazione di Pearson tra il campo *Esamina* e ogni campo correlato, nonché etichette descrittive facoltative per i valori di correlazione. Per ulteriori informazioni, consultare l'argomento "Impostazioni di correlazione" a pagina 294.

Menu Genera. Il menu Genera rende disponibili operazioni di generazione dei nodi.

- **Filtro.** Genera un nodo Filtro per filtrare i campi di output non correlati o con una correlazione debole con altri campi.

Generazione di un nodo Filtro da analisi statistiche

Il nodo Filtro generato da un browser di output del nodo Statistiche filtrerà i campi in base alle relative correlazioni con altri campi. Le correlazioni vengono ordinate per valore assoluto, utilizzando le correlazioni maggiori (in base ai criteri impostati nella finestra di dialogo Genera filtro da statistiche) e creando un filtro che passa tutti i campi che vengono visualizzati in ciascuna di tali correlazioni di grandi dimensioni.

Modalità. Specificare la modalità di selezione delle correlazioni. **Includi** consente di conservare i campi inclusi nelle correlazioni specificate. **Escludi** consente di filtrare tali campi.

Includi/Escludi i campi che compaiono in. Specificare il criterio per la selezione delle correlazioni.

- **Prime N correlazioni.** Seleziona il numero specificato di correlazioni e include/esclude i campi presenti in tali correlazioni.
- **Percentuale delle prime correlazioni (%).** Seleziona la percentuale specificata ($n\%$) di correlazioni e include/esclude i campi presenti in tali correlazioni.
- **Correlazioni maggiori di.** Seleziona le correlazioni maggiori in termini di valore assoluto rispetto alla soglia specificata.

Nodo Medie

Il nodo Medie confronta le medie tra gruppi indipendenti o coppie di campi correlati per verificare se esiste una differenza significativa. Per esempio, è possibile confrontare le entrate medie prima e dopo il lancio di una promozione, oppure confrontare le entrate determinate da clienti che non hanno ricevuto la promozione con quelli che l'hanno ricevuta.

Le medie possono essere confrontate in due diversi modi, a seconda dei dati disponibili:

- **Tra gruppi all'interno di un campo.** Per confrontare gruppi indipendenti, selezionare un campo di verifica e un campo di raggruppamento. Per esempio, si potrebbe escludere un campione di clienti "di holdout" dall'invio di una promozione e confrontare le entrate medie del gruppo di controllo con quelle di tutti gli altri gruppi. In questo caso si imposterebbe un solo campo di verifica che indica le entrate per ogni singolo cliente, con un campo flag o nominale che indica se il cliente ha ricevuto l'offerta. I campioni sono indipendenti nel senso che ogni record viene assegnato all'uno o all'altro gruppo e non esiste un modo per collegare un membro specifico di un gruppo a un membro specifico di un altro. È possibile anche specificare un campo nominale con più di due valori per confrontare le medie di più gruppi. Quando viene eseguito, il nodo calcola un'analisi della varianza (ANOVA) monodirezionale sui campi selezionati. Nei casi in cui vi siano solo due gruppi di campi, i risultati dell'analisi ANOVA monodirezionale sono essenzialmente uguali a quelli ottenuti con un t test per campioni indipendenti. Per ulteriori informazioni, consultare l'argomento "Confronto di medie per gruppi indipendenti".
- **Tra coppie di campi.** Quando si confrontano le medie di due campi correlati, è necessario che i gruppi siano in qualche modo accoppiati per ottenere risultati significativi. Per esempio, si potrebbero confrontare le entrate medie provenienti dallo stesso gruppo di clienti prima e dopo l'invio di una promozione, o confrontare le percentuali d'uso di un servizio tra coppie marito-moglie per verificare eventuali differenze. Ogni record contiene due misure distinte ma correlate che si possono confrontare in modo significativo. Quando viene eseguito, il nodo calcola un t test per campioni accoppiati su ogni coppia di campi selezionati. Per ulteriori informazioni, consultare l'argomento "Confronto delle medie tra campi accoppiati" a pagina 297.

Confronto di medie per gruppi indipendenti

Selezionare **Tra gruppi all'interno di un campo** nel nodo Medie per confrontare la media di due o più gruppi indipendenti.

Campo di raggruppamento. Selezionare un flag numerico o un campo nominale con due o più valori distinti che divide i record nei gruppi da confrontare, per esempio i soggetti che hanno ricevuto un'offerta e quelli che non l'hanno ricevuta. Indipendentemente dal numero dei campi di verifica, è possibile selezionare un solo campo di raggruppamento.

Verifica campo/i. Selezionare uno o più campi numerici contenenti le misure da verificare. Per ogni campo selezionato verrà eseguita una verifica separata. Per esempio, si potrebbe verificare l'impatto di una determinata promozione sull'uso, sulle entrate e sul tasso di abbandono.

Confronto delle medie tra campi accoppiati

Selezionare **Tra coppie di campi** nel nodo Medie per confrontare le medie tra campi separati. Affinché i risultati siano significativi, i campi devono essere in qualche modo correlati (per esempio, entrate prima e dopo una promozione). È possibile anche selezionare più coppie di campi.

Campo uno. Selezionare un campo numerico contenente la prima delle misure da confrontare. In un'indagine di tipo "prima e dopo", questo sarebbe il campo "prima".

Campo due. Selezionare il secondo campo da confrontare.

Aggiungi. Aggiunge la coppia selezionata all'elenco Verifica coppia/e di campi.

Se necessario, ripetere la selezione di campi per aggiungere più coppie all'elenco.

Impostazioni di correlazione. Consente di specificare le opzioni per l'assegnazione di etichette all'intensità delle correlazioni. Per ulteriori informazioni, consultare l'argomento "Impostazioni di correlazione" a pagina 294.

Opzioni del nodo Medie

La scheda Opzioni consente di impostare i valori p della soglia utilizzati per identificare i risultati come importanti, marginali o non importanti. È inoltre possibile modificare l'etichetta di ogni classificazione. L'importanza viene misurata su una scala percentuale e, in termini generici, può essere definita come 1 meno la probabilità di ottenere un risultato (quale la differenza tra le medie di due campi) altrettanto o più estremo di quello osservato determinato solo dal caso. Per esempio, un valore p superiore a 0,95 indica una probabilità inferiore al 5% che il risultato possa essere dovuto solamente al caso.

Etichette di importanza. È possibile modificare le etichette utilizzate per identificare ogni coppia o gruppo di campi nell'output. Le etichette di default sono *importante*, *marginale* e *non importante*.

Valori di interruzione. Specifica la soglia per ogni rango. In genere, i valori p maggiori di 0,95 classificano gli elementi come importanti, mentre quelli minori di 0,9 sono classificati come non importanti, ma queste soglie si possono modificare in funzione delle esigenze.

Nota: le misure di importanza sono disponibili in numerosi nodi. I calcoli specifici dipendono dal nodo e dal tipo di obiettivo e di campi di input utilizzati, ma i valori possono essere comunque confrontati in quanto sono tutti misurati in base a una scala percentuale.

Browser di output del nodo Medie

Il browser di output del nodo Medie visualizza i dati delle tavole di contingenza e consente di eseguire operazioni standard, tra cui la selezione e la copia della tabella una riga alla volta, l'ordinamento in base a qualsiasi colonna ed il salvataggio e la stampa della tabella. Per ulteriori informazioni, consultare l'argomento "Visualizzazione dell'output" a pagina 272.

Le informazioni specifiche contenute nella tabella dipendono dal tipo di confronto (gruppi all'interno di un campo o campi separati).

Ordina per. Consente di ordinare l'output in base ad una colonna specifica. Fare clic sulla freccia su o giù per modificare la direzione dell'ordinamento. In alternativa, è possibile fare clic sull'intestazione di qualsiasi colonna per ordinare in base a quella colonna. Fare clic nuovamente per cambiare la direzione dell'ordinamento all'interno della colonna.

Visualizza. È possibile selezionare **Livello base** o **Opzioni avanzate** per controllare il livello di dettaglio visualizzato. La visualizzazione con opzioni avanzate comprende tutte le informazioni contenute in quella di livello base, ma con una serie di dettagli aggiuntivi.

Output medie - Confronto tra gruppi all'interno di un campo

Quando si confrontano gruppi all'interno di un campo, il nome del campo di raggruppamento è visualizzato sopra la tabella di output e le medie e le relative statistiche sono riportate separatamente per ogni gruppo. La tabella comprende una riga separata per ogni campo di verifica.

Sono visualizzate le seguenti colonne:

- **Campo.** Elenca i nomi dei campi di verifica selezionati.
- **Medie di gruppo.** Visualizza la media per ogni categoria del campo di raggruppamento. Ad esempio, è possibile confrontare i soggetti che hanno ricevuto un'offerta speciale (*Nuova promozione*) con coloro che non l'hanno ricevuta (*Standard*). Nella visualizzazione con opzioni avanzate sono visualizzati anche la deviazione standard, l'errore standard e il conteggio.
- **Importanza.** Visualizza il valore e l'etichetta dell'importanza. Per ulteriori informazioni, consultare l'argomento "Opzioni del nodo Medie" a pagina 297.

Output avanzato

La visualizzazione con opzioni avanzate contiene le seguenti colonne aggiuntive.

- **F-Test.** Questo test si basa sul rapporto tra varianza fra i gruppi e varianza all'interno di ogni gruppo. Se le medie sono uguali per tutti i gruppi, il rapporto *F* dovrebbe avvicinarsi a 1 poiché entrambi i valori sono stime della varianza della stessa popolazione. Maggiore è il rapporto, maggiori sono la variazione tra i gruppi e la probabilità che esista una differenza significativa.
- **df.** Visualizza i gradi di libertà.

Output Medie - Confronto tra coppie di campi

Quando si confrontano campi separati, la tabella di output comprende una riga per ogni coppia di campi selezionata.

- **Campo uno/due.** Visualizza il nome del primo e del secondo campo di ogni coppia. Nella visualizzazione con opzioni avanzate sono visualizzati anche la deviazione standard, l'errore standard e il conteggio.
- **Media uno/due.** Visualizza la media rispettiva dei due campi.
- **Correlazione.** Misura l'intensità della relazione tra due campi continui (intervallo numerico). I valori vicini a +1.0 indicano una forte associazione positiva ed i valori vicini a -1.0 indicano una forte associazione negativa. Per ulteriori informazioni, consultare l'argomento "Impostazioni di correlazione" a pagina 294.
- **Differenza media.** Visualizza la differenza tra le medie dei due campi.
- **Importanza.** Visualizza il valore e l'etichetta dell'importanza. Per ulteriori informazioni, consultare l'argomento "Opzioni del nodo Medie" a pagina 297.

Output avanzato

L'opzione Output avanzato aggiunge le seguenti colonne:

Intervallo di confidenza del 95%. Limiti inferiore e superiore dell'intervallo all'interno del quale è probabile che si trovi la media vera nel 95% di tutti i possibili campioni di quelle dimensioni e di quella popolazione.

T-Test. La statistica *t* si ottiene dividendo la differenza media per il suo errore standard. Maggiore è il valore assoluto di questa statistica, maggiore è la probabilità che le medie non siano uguali.

df. Visualizza i gradi di libertà della statistica.

Nodo Report

Il nodo Report consente di creare report formattati che contengono testo fisso e dati ed altre espressioni derivate dai dati. Il formato del report viene specificato utilizzando modelli di testo per definire il testo fisso e le costruzioni di output dei dati. È possibile fornire una formattazione personalizzata del testo utilizzando tag HTML nel modello e impostando apposite opzioni nella scheda Output. I valori dei dati e il restante output condizionale vengono inclusi nel report utilizzando espressioni CLEM nel modello.

Alternative al nodo Report

Il nodo Report viene soprattutto utilizzato per elencare i record o i casi prodotti da un flusso, per esempio tutti i record che soddisfano una determinata condizione. In questo senso, può essere considerato come un'alternativa meno strutturata al nodo Tabella.

- Se si desidera produrre un report che, invece dei dati veri e propri, elenchi informazioni sui campi o altri elementi definiti nel flusso (per esempio le definizioni di campi specificate in un nodo Tipo), è possibile utilizzare uno script.
- Per generare un report che includa più oggetti di output (quali una raccolta di modelli, tabelle e grafici generati da uno o più flussi) e che possano essere prodotti in più formati (come testo, HTML e Microsoft Word/Office) è possibile utilizzare un progetto IBM SPSS Modeler.
- Per generare un elenco di nomi di campi senza utilizzare la funzionalità di script, è possibile utilizzare un nodo Tabella preceduto da un nodo Campione che scarta tutti i record. In questo modo viene generata una tabella senza righe, che può essere traspunta nelle esportazioni per creare un elenco di nomi di campi in un'unica colonna (a questo scopo, selezionare **Trasponi dati** nella scheda Output del nodo Tabella).

Scheda Modello del nodo Report

Creazione di un modello. Per definire il contenuto del report, generare un modello nella scheda Modello del nodo Report. Il modello è formato da righe di testo, ognuna delle quali contiene una specifica relativa al contenuto del report, e da speciali righe di tag utilizzate per indicare l'ambito delle righe del contenuto. In ogni riga del contenuto, le espressioni CLEM racchiuse tra parentesi quadre ([]) vengono valutate prima dell'invio della riga al report. Per una riga del modello esistono tre possibili ambiti:

Fisso. Le righe che non hanno contrassegni di altro genere sono considerate fisse. Le righe fisse vengono copiate nel report solo una volta, dopo la valutazione delle espressioni in esse contenute. Per esempio, la riga

```
Questo report è stato generato [@TODAY]
```

copierà nel report una singola riga, contenente il testo e la data corrente.

Globale (iterazione completa). Le righe tra i tag speciali #ALL e # vengono copiate nel report una volta per ogni record di dati di input. Le espressioni CLEM (racchiuse tra parentesi) vengono valutate in base al record corrente per ciascuna riga di output. Per esempio, le righe

```
#ALL  
For record [@INDEX], the value of AGE is [AGE]  
#
```

includono una riga per ogni record che indichi il numero di record e l'età.

Per generare un elenco di tutti i record:

```
#ALL  
[Age] [Sex] [Cholesterol] [BP]  
#
```

Condizionale (iterazione condizionale). Le righe tra i tag speciali #WHERE <condizione> e # vengono copiate nel report una volta per ogni record in cui la condizione specificata è vera. La condizione è un'espressione CLEM. Nella condizione WHERE, le parentesi sono facoltative. Per esempio, le righe

```
#WHERE [SEX = 'M']
Male at record no. [@INDEX] has age [AGE].
#
```

scrivono nel file una riga per ogni record con valore *M* per il sesso. Il report completo conterrà le righe fisse, globali e condizionali definite applicando il modello ai dati di input.

È possibile specificare le opzioni di visualizzazione o salvataggio dei risultati mediante la scheda Output, comune a diversi tipi di nodi Output. Per ulteriori informazioni, consultare l'argomento "Scheda Output del nodo Output" a pagina 276.

Output di dati in formato HTML o XML

I tag HTML o XML possono essere inclusi direttamente nel modello per scrivere report in uno di questi due formati. Per esempio, il seguente modello genera una tabella HTML.

```
This report is written in HTML.
Only records where Age is above 60 are included.
```

```
<HTML>
  <TABLE border="2">
    <TR>
      <TD>Age</TD>
      <TD>BP</TD>
      <TD>Cholesterol</TD>
      <TD>Drug</TD>
    </TR>

    #WHERE Age > 60
    <TR>
      <TD>[Age]</TD>
      <TD>[BP]</TD>
      <TD>[Cholesterol]</TD>
      <TD>[Drug]</TD>
    </TR>
  #
</TABLE>
</HTML>
```

Browser di output del nodo Report

Il browser del nodo Report mostra il contenuto del report generato. Le normali opzioni di salvataggio, esportazione e stampa sono disponibili dal menu File e le normali opzioni di modifica sono disponibili dal menu Modifica. Per ulteriori informazioni, consultare l'argomento "Visualizzazione dell'output" a pagina 272.

Nodo Calcola globali

Il nodo Calcola globali analizza i dati e calcola i valori di riepilogo che possono essere utilizzati nelle espressioni CLEM. Per esempio, è possibile utilizzare un nodo Calcola globali per calcolare le statistiche di un campo denominato *età* e utilizzare quindi la media globale dell'*età* nelle espressioni CLEM inserendo la funzione @GLOBAL_MEAN(età).

Scheda Impostazioni del nodo Calcola globali

Valori globali da creare. Selezionare il campo o i campi per i quali si desidera che siano disponibili valori globali. È possibile selezionare più campi. Per ogni campo specificare le statistiche da calcolare, accertandosi che siano selezionate nelle colonne accanto al nome del campo.

- **MEAN.** Valore medio (media) del campo in tutti i record.
- **SUM.** Somma dei valori del campo in tutti i record.
- **MIN.** Valore minimo del campo.
- **MAX.** Valore massimo del campo.
- **SDEV.** Deviazione standard, che misura la variabilità dei valori di un campo ed è calcolata come radice quadrata della varianza.

Operazioni di default. Le opzioni qui selezionate verranno utilizzate quando verranno aggiunti nuovi campi all'elenco Valori globali sopra riportato. Per modificare l'insieme di default delle statistiche, selezionare o deselezionare le statistiche come desiderato. È inoltre possibile utilizzare il pulsante **Applica** per applicare le operazioni di default a tutti i campi compresi nell'elenco.

Cancella tutti i valori globali prima dell'esecuzione. Selezionare questa opzione per rimuovere tutti i valori globali prima di calcolare nuovi valori. Se questa opzione non è selezionata, i nuovi valori calcolati sostituiranno i precedenti, tuttavia i valori globali non ricalcolati continueranno a essere disponibili.

Visualizza anteprema dei valori globali creati dopo l'esecuzione. Se si seleziona questa opzione, dopo l'esecuzione verrà visualizzata la scheda Valori globali della finestra di dialogo delle proprietà del flusso contenente i valori globali calcolati.

Nodo Adattamento simulazione

Il nodo Adattamento simulazione adatta una serie di distribuzioni statistiche candidate a ciascun campo dei dati. L'adattamento di ciascuna distribuzione ad un campo viene valutato utilizzando un criterio di bontà di adattamento. Quando viene eseguito un nodo Adattamento simulazione, viene creato un nodo Genera simulazione (o viene aggiornato un nodo esistente). Ad ogni campo viene assegnata la distribuzione con l'adattamento migliore. Il nodo Genera simulazione può quindi essere utilizzato per generare i dati simulati per ciascun campo.

Sebbene il nodo Adattamento simulazione sia un nodo terminale, non aggiunge un modello alla tavolozza dei modelli generati, aggiunge un output o un grafico alla scheda degli output, o esporta i dati.

Nota: Se i dati cronologici sono scarsi (ossia mancano molti valori) potrebbe essere difficile per il componente di adattamento individuare sufficienti valori validi per adattare le distribuzioni ai dati. Per adattare le distribuzioni ai dati, il componente di adattamento necessita di 2000 valori validi. Nei casi in cui i dati sono scarsi, prima di eseguire l'adattamento rimuovere i campi privi di valori, se non sono necessari, o assegnare i valori mancanti. Utilizzando le opzioni della scheda **Qualità** del nodo Verifica dati è possibile visualizzare il numero di record completi, identificare quali sono i campi con valori mancanti e selezionare un metodo di assegnazione. Se non vi è un numero sufficiente di record per l'adattamento delle distribuzioni, è possibile utilizzare un nodo di bilanciamento per aumentare il numero di record.

Utilizzo di un nodo Adattamento simulazione per creare automaticamente un nodo Genera simulazione

La prima volta che il nodo Adattamento simulazione viene eseguito, viene creato un nodo Genera simulazione con un collegamento di aggiornamento al nodo Adattamento simulazione. Se il nodo Adattamento simulazione viene eseguito nuovamente, viene creato un nuovo nodo Genera simulazione solo se il collegamento di aggiornamento è stato rimosso. Un nodo Adattamento simulazione può inoltre essere utilizzato per aggiornare un nodo Genera simulazione connesso. Il risultato dipende dalla presenza

o meno degli stessi campi in entrambi i nodi e se i campi sono sbloccati nel nodo Genera simulazione. Per ulteriori informazioni, consultare l'argomento "Nodo Genera simulazione" a pagina 42.

Un nodo Adattamento simulazione può avere un solo collegamento di aggiornamento ad un nodo Genera simulazione. Per definire un collegamento di aggiornamento ad un nodo Genera simulazione, effettuare le seguenti operazioni:

1. Fare clic con il tasto destro del mouse sul nodo Adattamento simulazione.
2. Dal menu, selezionare **Definisci collegamento di aggiornamento**.
3. Fare clic sul nodo Genera simulazione al quale si desidera definire un collegamento di aggiornamento.

Per rimuovere un collegamento di aggiornamento tra un nodo Adattamento simulazione e un nodo Genera simulazione, fare clic con il tasto destro del mouse sul collegamento di aggiornamento e selezionare **Rimuovi collegamento**.

Adattamento delle distribuzioni

Una distribuzione statistica è la frequenza teorica delle occorrenze dei valori che una variabile può assumere. Nel nodo Adattamento simulazione, un insieme di distribuzioni statistiche teoriche viene confrontato con ciascun campo di dati. Le distribuzioni disponibili per l'adattamento sono descritte nell'argomento "Distribuzioni" a pagina 53. I parametri della distribuzione teorica vengono adeguati per fornire il miglior adattamento ai dati secondo una misurazione della bontà di adattamento; il criterio Anderson-Darling o il criterio Kolmogorov-Smirnov. I risultati dell'adattamento delle distribuzioni in base al nodo Adattamento simulazione mostrano quali distribuzioni sono adatte, le migliori stime dei parametri per ciascuna distribuzione e in che misura la distribuzione si adatta ai dati. Durante l'adattamento delle distribuzioni, vengono calcolate anche le correlazioni tra i campi con tipi di archiviazione numerica e le contingenze tra i campi con distribuzione di categoria. I risultati dell'adattamento delle distribuzioni vengono utilizzati per creare un nodo Genera simulazione.

Prima di adattare qualsiasi distribuzione ai dati, vengono esaminati i primi 1000 record per la presenza di valori mancanti. Se sono presenti troppi valori mancanti, l'adattamento delle distribuzioni non è possibile. In questo caso, è necessario decidere se una delle seguenti opzioni è appropriata:

- Utilizzare un nodo upstream per rimuovere i record con i valori mancanti.
- Utilizzare un nodo upstream per assegnare i valori ai valori mancanti.

L'adattamento delle distribuzioni non esclude i valori mancanti dell'utente. Se i dati contengono valori mancanti dell'utente e si desidera che quei valori vengano esclusi dall'adattamento delle distribuzioni, impostare tali valori su mancanti di sistema.

Il ruolo di un campo non viene preso in considerazione quando le distribuzioni vengono adattate. Ad esempio, i campi con ruolo **Target** vengono trattati allo stesso modo dei campi con i ruoli **Input**, **Nessuno**, **Entrambi**, **Partizione**, **Suddivisione**, **Frequenza** e **ID**.

I campi vengono trattati in modo diverso durante l'adattamento delle distribuzioni in base al tipo di archiviazione e al livello di misurazione. Il trattamento dei campi durante l'adattamento delle distribuzioni è descritto nella tabella riportata di seguito.

Tabella 51. Adattamento delle distribuzioni in base al tipo di archiviazione e al livello di misurazione dei campi

Tipo di archiviazione	Livello di misurazione					
	Continuo	Categoriale	Indicatore	Nominale	Ordinale	Senza tipo
Stringa	Impossibile		Vengono adattate le distribuzioni di categoria, cubo e fisse			

Tabella 51. Adattamento delle distribuzioni in base al tipo di archiviazione e al livello di misurazione dei campi (Continua)

Tipo di archiviazione	Livello di misurazione			
Intero	Vengono adattate tutte le distribuzioni. Vengono calcolate le correlazioni e le contingenze.	Viene adattata la distribuzione di categoria. Le correlazioni non vengono calcolate.	Vengono adattate le distribuzioni binomiale, binomiale negativa e Poisson, e le correlazioni vengono calcolate.	Il campo viene ignorato e non viene trasmesso al nodo Genera simulazione.
Reale				
Ora				
Data				
Timestamp				
Sconosciuto	Il tipo di archiviazione appropriato viene determinato dai dati.			

I campi con livello di misurazione ordinale vengono trattati come campi continui e vengono inclusi nella tabella di correlazioni nel nodo Genera simulazione. Se si desidera che una distribuzione diversa da binomiale, binomiale negativa o Poisson venga adattata ad un campo ordinale, è necessario modificare il livello di misurazione del campo su continuo. Se in precedenza è stata definita un'etichetta per ciascun valore di campo ordinale e si modifica il livello di misurazione su continuo, le etichette andranno perse.

I campi che hanno valori singoli non vengono trattati in modo diverso durante l'adattamento delle distribuzioni rispetto ai campi con valori multipli. I campi con tipo di archiviazione ora, data o timestamp vengono trattati come numerici.

Adattamento delle distribuzioni ai campi di suddivisione

Se i dati contengono un campo di suddivisione e si desidera eseguire l'adattamento delle distribuzioni separatamente per ciascuna suddivisione, è necessario trasformare i dati utilizzando un nodo upstream Riorganizza. Utilizzando il nodo Riorganizza, generare un nuovo campo per ogni valore del campo di suddivisione. Questi dati riorganizzati possono quindi essere utilizzati per l'adattamento delle distribuzioni nel nodo Adattamento simulazione.

Scheda Impostazioni del nodo Adattamento simulazione

Nome nodo Origine. È possibile generare automaticamente il nome del nodo Genera simulazione generato (o aggiornato) selezionando **Automatico**. Il nome generato automaticamente è il nome specificato nel nodo Adattamento simulazione se è stato specificato un nome personalizzato (o Sim Gen se non è stato specificato un nome personalizzato nel nodo Adattamento simulazione). Selezionare **Personalizzato** per specificare un nome personalizzato nel campo di testo adiacente. A meno che non venga modificato il campo di testo, il nome personalizzato di default è Sim Gen.

Opzioni di adattamento Con queste opzioni è possibile specificare in che modo le distribuzioni vengono adattate ai campi e come viene valutato l'adattamento delle distribuzioni.

- **Numero di casi da campionare.** Specifica il numero di casi da utilizzare durante l'adattamento delle distribuzioni ai campi nel dataset. Selezionare **Tutti i casi** per adattare le distribuzioni a tutti i record contenuti nei dati. Se il dataset è di grandi dimensioni, si potrebbe voler limitare il numero di casi utilizzati per l'adattamento delle distribuzioni. Selezionare **Limita ai primi N casi** per utilizzare solo i primi N casi. Fare clic sulle frecce per specificare il numero di casi da utilizzare. In alternativa, è possibile utilizzare un nodo upstream per prendere un campione casuale di record per l'adattamento delle distribuzioni.
- **Criteri bontà dell'adattamento (solo campi continui).** Per i campi continui, selezionare il test Anderson-Darling o il test Kolmogorov-Smirnoff di bontà dell'adattamento per classificare le distribuzioni durante l'adattamento delle distribuzioni ai campi. Il test Anderson-Darling è selezionato di default ed è consigliato specialmente quando si desidera assicurare il miglior adattamento possibile nelle aree di coda. Entrambe le statistiche vengono calcolate per ogni distribuzione candidata ma solo la statistica selezionata viene utilizzata per ordinare le distribuzioni e determinare quella con l'adattamento migliore.
- **Bin (solo distribuzione empirica).** Per i campi continui, la distribuzione empirica è la funzione di distribuzione cumulativa dei dati cronologici. È la probabilità di ciascun valore o intervallo di valori, e viene ricavata direttamente dai dati. È possibile specificare il numero di bin utilizzati per il calcolo della distribuzione empirica per i campi continui facendo clic sulle frecce. Il valore di default è 100 e il valore massimo è 1000.
- **Campo peso (facoltativo).** Se il dataset contiene un campo peso, fare clic sull'icona di selezione del campo e selezionare il campo peso dall'elenco. Il campo peso viene quindi escluso dal processo di adattamento delle distribuzioni. L'elenco presenta tutti i campi del dataset con livello di misurazione continuo. È possibile selezionare un solo campo peso.

Nodo Valutazione simulazione

Il nodo Valutazione simulazione è un nodo terminale che valuta un campo specificato, fornisce una distribuzione del campo e produce i grafici delle distribuzioni e delle correlazioni. Questo nodo viene utilizzato principalmente per valutare i campi continui. Pertanto completa il grafico di valutazione, generato da un nodo valutazione, ed è utile per la valutazione dei campi discreti. Un'altra differenza è che il nodo Valutazione simulazione valuta una singola previsione in diverse iterazioni, mentre il nodo Valutazione valuta più previsioni ognuna con una singola iterazione. Le iterazioni vengono generate quando viene specificato più di un valore per un parametro di distribuzione nel nodo Genera simulazione. Per ulteriori informazioni, consultare l'argomento "Iterazioni" a pagina 52.

Il nodo Valutazione simulazione è progettato per essere utilizzato con dati ottenuti dai nodi Adattamento simulazione e Genera simulazione. Il nodo può, tuttavia, essere utilizzato con qualsiasi altro nodo. È possibile inserire qualsiasi numero di fasi di elaborazione tra il nodo Genera simulazione e il nodo Valutazione simulazione.

Importante: Il nodo Valutazione simulazione richiede almeno 1000 record con valori validi per il campo obiettivo.

Scheda Impostazioni del nodo Valutazione simulazione

Nella scheda Impostazioni del nodo Valutazione simulazione, è possibile specificare il ruolo di ciascun campo nel dataset e personalizzare l'output generato dalla simulazione.

Seleziona un elemento. Consente di cambiare vista tra le tre disponibili per il nodo Valutazione simulazione: Campi, Funzioni di densità e Output.

Vista Campi

Campo Obiettivo. Si tratta di un campo obbligatorio. Fare clic su sulla freccia per selezionare il campo obiettivo del dataset dall'elenco a discesa. Il campo selezionato può avere un livello di misurazione continuo, ordinale o nominale ma non può avere la data o un livello di misurazione non specificato.

Campo Iterazione (facoltativo). Se i dati hanno un campo iterazione che indica a quale iterazione appartiene ciascun record nei dati, è necessario selezionarlo qui. Questo significa che ogni iterazione verrà valutata separatamente. Solo i campi con i livelli di misurazione continuo, ordinale o nominale possono essere selezionati.

I dati di input sono già ordinati dall'iterazione. È abilitato solo se è specificato un campo di iterazione nel campo **Campo Iterazione (facoltativo)**. Selezionare questa opzione solo se si è certi che i dati di input sono già ordinati dal campo iterazione specificato in **Campo Iterazione (facoltativo)**.

Numero massimo di iterazioni da tracciare. È abilitato solo se è specificato un campo di iterazione nel campo **Campo Iterazione (facoltativo)**. Fare clic sulle frecce per specificare il numero di iterazioni da tracciare. Specificando questo numero si evita di tracciare troppe iterazioni su un singolo grafico, che ne renderebbe difficile l'interpretazione. Il livello più basso a cui il numero massimo di iterazioni può essere impostato è 2; il livello più alto è 50. Il numero massimo di iterazioni da tracciare inizialmente è impostato su 10.

Campi di input per tornado di correlazione. Il grafico tornado di correlazione è un grafico a barre che visualizza i coefficienti di correlazione tra l'obiettivo specificato e ognuno degli input specificati. Fare clic sull'icona di selezione del campo per selezionare i campi di input da includere nel grafico tornado, da un elenco degli input simulati disponibili. Solo i campi con i livelli di misurazione continuo e ordinale possono essere selezionati. I campi di input nominale, di data e senza tipo non sono disponibili nell'elenco e non possono essere selezionati.

Vista Funzioni di densità

Con le opzioni contenute in questa vista, è possibile personalizzare l'output per le funzioni di densità di probabilità e le funzioni di distribuzione cumulativa per target continui, ed anche i grafici a barre di valori previsti per target di categoria.

Funzioni di densità. Le funzioni di densità sono lo strumento principale per analizzare l'insieme di risultati ottenuti dalla simulazione.

- **Funzione di densità di probabilità (PDF).** Selezionare questa opzione per produrre una funzione di densità di probabilità per il campo obiettivo. La funzione di densità di probabilità visualizza la distribuzione dei valori obiettivo. È possibile utilizzare la funzione di densità di probabilità per determinare la probabilità che il target sia all'interno di una particolare area. Per i target di categoria (target con un livello di misurazione nominale o ordinale) viene generato un grafico a barre che visualizza la percentuale di casi che rientrano in ciascuna categoria del target.
- **Funzione di distribuzione cumulativa (CDF).** Selezionare questa opzione per produrre una funzione di distribuzione cumulativa per il campo obiettivo. La funzione di distribuzione cumulativa visualizza la probabilità che il valore del target sia inferiore o uguale ad un valore specificato. È disponibile solo per target continui.

Linee di riferimento (continue). Queste opzioni sono abilitate se la **Funzione di densità di probabilità (PDF)**, la **Funzione di distribuzione cumulativa (CDF)** o entrambe sono selezionate. Con queste opzioni è possibile aggiungere varie linee di riferimento fisse verticali alle funzioni di densità di probabilità e alle funzioni di distribuzione cumulativa.

- **Media.** Selezionare questa opzione per aggiungere una linea di riferimento sul valore medio del campo obiettivo.
- **Mediana.** Selezionare questa opzione per aggiungere una linea di riferimento sul valore mediano del campo obiettivo.
- **Deviazioni standard.** Selezionare questa opzione per aggiungere una linea di riferimento sul più e meno un numero specificato di deviazioni standard dalla media del campo obiettivo. La selezione di questa opzione abilita il campo contiguo **Numero**. Fare clic sulle frecce per specificare il numero di deviazioni standard. Il numero minimo di deviazioni standard è 1 e il numero massimo è 10. Il numero di deviazioni standard viene inizialmente impostato su 3.

- **Percentili.** Selezionare questa opzione per aggiungere le linee di riferimento su due valori di percentile della distribuzione del campo obiettivo. La selezione di questa opzione abilita i campi contigui di testo **Inferiore** e **Superiore**. Ad esempio, l'immissione del valore 90 nel campo di testo **Superiore** aggiunge una linea di riferimento al 90° percentile del target, che è il valore sotto il quale rientra il 90% delle osservazioni. Allo stesso modo, il valore 10 nel campo di testo **Inferiore** rappresenta il 10° percentile del target, che è il valore sotto il quale rientra il 10% delle osservazioni.
- **Righe di riferimento personalizzate.** Selezionare questa opzione per aggiungere le righe di riferimento nei valori specificati lungo l'asse orizzontale. La selezione di questa opzione abilita la tabella contigua **Valori**. Ogni volta che si immette un numero valido nella tabella **Valori**, una nuova riga vuota viene accodata in fondo alla tabella. Un numero *valido* è un numero compreso nell'intervallo di valori del campo obiettivo

Nota: Quando più funzioni di densità o di distribuzione (da più iterazioni) vengono visualizzate su un singolo grafico, le linee di riferimento (diverse dalle linee personalizzate) vengono applicate separatamente a ciascuna funzione.

Target di categoria (solo PDF). Queste opzioni sono abilitate solo se la **Funzione di densità di probabilità (PDF)** è selezionata.

- **Valori di categoria da segnalare.** Per i modelli con campi obiettivo di categoria, il risultato del modello è un insieme di probabilità previste, una per ogni categoria, che il valore obiettivo rientri in ciascuna categoria. La categoria con la probabilità più elevata viene considerata come categoria prevista ed utilizzata per generare il grafico a barre per la funzione di densità di probabilità. Selezionare **Categoria prevista** per generare il grafico a barre. Selezionare **Probabilità previste** per generare istogrammi della distribuzione delle probabilità previste per ognuna delle categorie del campo obiettivo. È possibile inoltre selezionare **Entrambi** per generare entrambi i tipi di grafico.
- **Raggruppamento per analisi di sensibilità.** Le simulazioni che includono iterazioni di analisi della sensibilità generano un campo obiettivo indipendente (o campo obiettivo previsto da un modello) per ogni iterazione definita dall'analisi. È presente un'iterazione per ogni valore del parametro di distribuzione variato. Quando sono presenti iterazioni, il grafico a barre della categoria prevista per un campo obiettivo di categoria viene visualizzato come grafico a barre con cluster che include i risultati di tutte le iterazioni. Selezionare **Raggruppa insieme categorie** o **Raggruppa insieme iterazioni**.

Vista Output

Valori percentile delle distribuzioni obiettivo. Con queste opzioni, è possibile scegliere di creare una tabella di valori percentile delle distribuzioni obiettivo e specificare i percentili da visualizzare.

Crea una tabella dei valori percentili. Per i campi di target continuo selezionare questa opzione per ottenere una tabella dei percentili specificati delle distribuzioni obiettivo. Scegliere una delle seguenti opzioni per specificare i percentili:

- **Quartili.** I quartili sono il 25esimo, 50esimo e 75esimo percentile della distribuzione di campi obiettivo. Le osservazioni sono suddivise in quattro gruppi di uguale dimensione.
- **Intervalli.** Se si desidera un numero uguale di gruppi diverso da quattro, selezionare **Intervalli**. La selezione di questa opzione abilita il campo contiguo **Numero**. Fare clic sulle frecce per specificare il numero di intervalli. Il numero minimo di intervalli è 2 e quello massimo è 100. Il numero di intervalli inizialmente è impostato su 10.
- **Percentili personalizzati.** Selezionare **Percentili personalizzati** per specificare i singoli percentili, ad esempio il 99esimo percentile. La selezione di questa opzione abilita la tabella contigua **Valori**. Ogni volta che si immette un numero valido, compreso tra 1 e 100, nella tabella **Valori**, una nuova riga vuota viene accodata in fondo alla tabella.

Output del nodo Valutazione simulazione

Quando viene eseguito il nodo Valutazione simulazione, l'output viene aggiunto al gestore di output. Il browser dell'output di Valutazione simulazione visualizza i risultati dell'esecuzione del nodo Valutazione

simulazione. Le normali opzioni di salvataggio, esportazione e stampa sono disponibili dal menu **File** e le normali opzioni di modifica sono disponibili dal menu **Modifica**. Per ulteriori informazioni, consultare l'argomento "Visualizzazione dell'output" a pagina 272. Il menu **Visualizza** è abilitato solo se è selezionato uno dei grafici. Non è abilitato per la tabella di distribuzione o gli output di informazioni. Dal menu **Visualizza** è possibile selezionare **Modalità Modifica** per modificare il layout e l'aspetto del grafico, oppure **Modalità Esplora** per esplorare i dati ed i valori rappresentati dal grafico. La modalità Statico fissa le linee di riferimento del grafico (e i dispositivi di scorrimento) nelle rispettive posizioni correnti in modo da non consentirne lo spostamento. La modalità Statico è l'unica modalità in cui è possibile copiare, esportare o stampare il grafico con le relative linee di riferimento. Per selezionare questa modalità, fare clic su **Modalità Statico** nel menu **Visualizza**.

La finestra del browser di output di Valutazione simulazione è composta da due pannelli. A sinistra nella finestra è presente un pannello di navigazione che visualizza le rappresentazioni in miniatura dei grafici generati durante l'esecuzione del nodo Valutazione simulazione. Quando una miniatura è selezionata, l'output del grafico viene visualizzato sul pannello a destra nella finestra.

Pannello di navigazione

Il pannello di navigazione del browser di output contiene le miniature dei grafici generati da una simulazione. Le miniature visualizzate sul pannello di navigazione dipendono dal livello di misurazione del campo obiettivo e dalle opzioni selezionate nella finestra di dialogo del nodo Valutazione simulazione. Le descrizioni delle miniature vengono fornite nella tabella riportata di seguito.

Tabella 52. Miniature del pannello di navigazione

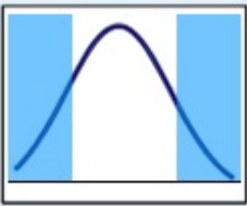
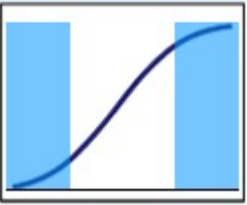
Miniatura	Descrizione	Commenti
	Funzione di densità di probabilità	Questa miniatura viene visualizzata solo se il livello di misurazione del campo obiettivo è continuo e Funzione di densità di probabilità (PDF) è selezionato nella vista Funzioni di densità della finestra di dialogo del nodo Valutazione simulazione. Se il livello di misurazione del campo obiettivo è di categoria, questa miniatura non viene visualizzata.
	Funzione di distribuzione cumulativa	Questa miniatura viene visualizzata solo se il livello di misurazione del campo obiettivo è continuo e Funzione di distribuzione cumulativa (CDF) è selezionato nella vista Funzioni di densità della finestra di dialogo del nodo Valutazione simulazione. Se il livello di misurazione del campo obiettivo è di categoria, questa miniatura non viene visualizzata.

Tabella 52. Miniature del pannello di navigazione (Continua)


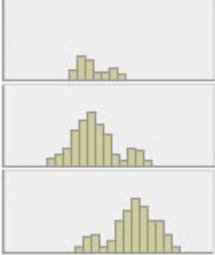
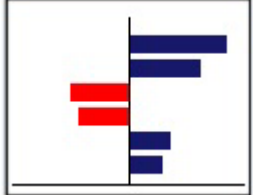
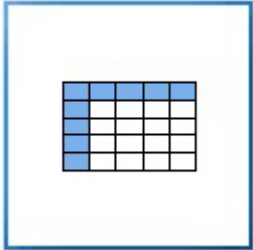

Miniatura	Descrizione	Commenti
	Valori di categoria previsti	<p>Questa miniatura viene visualizzata solo se il livello di misurazione del campo obiettivo è di categoria, Funzione di densità di probabilità (PDF) è selezionato nella vista Funzioni di densità della finestra di dialogo del nodo Valutazione simulazione e Categoria prevista o Entrambi è selezionato nell'area Valori di categoria da riportare.</p> <p>Se il livello di misurazione del campo obiettivo è continuo, questa miniatura non viene visualizzata.</p>
	Probabilità di categoria previste	<p>Questa miniatura viene visualizzata solo se il livello di misurazione del campo obiettivo è di categoria, Funzione di densità di probabilità (PDF) è selezionato nella vista Funzioni di densità della finestra di dialogo del nodo Valutazione simulazione e Probabilità previste o Entrambi è selezionato nell'area Valori di categoria da riportare.</p> <p>Se il livello di misurazione del campo obiettivo è continuo, questa miniatura non viene visualizzata.</p>
	Grafici tornado	<p>Questa miniatura viene visualizzata solo se uno o più campi di input sono selezionati nel campo Campi di input per tornado di correlazione nella vista Campi della finestra di dialogo del nodo Valutazione simulazione.</p>
	Tabella di distribuzione	<p>Questa miniatura viene visualizzata solo se il livello di misurazione del campo obiettivo è continuo e Crea una tabella di valori percentile è selezionato nella vista Output della finestra di dialogo del nodo Valutazione simulazione. Il menu Visualizza è disabilitato per questo grafico.</p> <p>Se il livello di misurazione del campo obiettivo è di categoria, questa miniatura non viene visualizzata.</p>

Tabella 52. Miniature del pannello di navigazione (Continua)

Miniatura	Descrizione	Commenti
	Informazioni	Questa miniatura viene sempre visualizzata. Il menu Visualizza è disabilitato per questo output.

Output dei grafici

I tipi di grafici di output disponibili dipendono dal livello di misurazione del campo obiettivo, se viene utilizzato un campo iterazione, e dalle opzioni selezionate nella finestra di dialogo del nodo Valutazione simulazione. Una serie di grafici generati da una simulazione dispongono di funzioni interattive che l'utente può utilizzare per personalizzare la visualizzazione. Le funzioni interattive sono disponibili facendo clic su **Opzioni grafico**. Tutti i grafici di simulazione sono visualizzazioni di tipo lavagna grafica.

Grafici Funzione di densità di probabilità per target continui. Questo grafico mostra sia la probabilità che la frequenza, con la scala di probabilità sull'asse verticale sinistro e la scala di frequenza sull'asse verticale destro. Il grafico dispone di due linee di riferimento scorrevoli verticali che dividono il grafico in aree separate. La tabella riportata sotto il grafico visualizza la percentuale di distribuzione presente in ognuna delle aree. Se sullo stesso grafico vengono visualizzate più funzioni di densità (a causa di iterazioni), la tabella dispone di una riga separata per le probabilità associate a ciascuna funzione di densità, e una colonna aggiuntiva che contiene il nome iterazione ed un colore associato a ciascuna funzione di densità. Le iterazioni sono elencate nella tabella in ordine alfabetico, in base all'etichetta dell'iterazione. Se non sono disponibili etichette dell'iterazione, viene utilizzato il valore dell'iterazione. Non è possibile modificare la tabella.

Ogni linea di riferimento dispone di un dispositivo di scorrimento (un triangolo invertito) che è possibile utilizzare per spostare facilmente la linea. Ogni dispositivo di scorrimento ha un'etichetta che ne indica la posizione corrente. Per default, i dispositivi di scorrimento sono posizionati al 5° e al 95° percentile della distribuzione. Se sono presenti più iterazioni, i dispositivi di scorrimento sono posizionati al 5° e al 95° percentile della prima iterazione elencata nella tabella. Non è possibile spostare le linee in modo che si incrocino le une con le altre.

Facendo clic su **Opzioni grafico** sono disponibili una serie di funzioni aggiuntive. In particolare, è possibile impostare esplicitamente le posizioni dei dispositivi di scorrimento, aggiungere linee di riferimento fisse e modificare la vista grafico da curva continua ad istogramma. Per ulteriori informazioni, consultare l'argomento "Grafici: opzioni" a pagina 311. Fare clic con il tasto destro del mouse sul grafico per copiare o esportare il grafico.

Grafici Funzione di distribuzione cumulativa per target continui. Questo grafico dispone delle stesse due linee di riferimento verticali spostabili e della tabella associata descritte per il grafico della funzione di densità di probabilità. I controlli dei dispositivi di scorrimento e la tabella funzionano allo stesso modo della funzione di densità di probabilità e presentano più iterazioni. Per le funzioni di distribuzione vengono utilizzati gli stessi colori utilizzati per identificare quale funzione di densità appartiene a ciascuna iterazione.

Questo grafico fornisce inoltre l'accesso alla finestra di dialogo Opzioni grafico, che consente di impostare esplicitamente le posizioni dei dispositivi di scorrimento, aggiungere linee di riferimento fisse e specificare se la funzione di distribuzione cumulativa viene visualizzata come funzione crescente (default) o come funzione decrescente. Per ulteriori informazioni, consultare l'argomento "Grafici: opzioni" a pagina 311

pagina 311. Fare clic con il tasto destro del mouse sul grafico per copiare, esportare o modificare il grafico. Selezionando **Modifica** il grafico viene aperto in una finestra mobile di editor con lavagna grafica.

Grafico Valori categoria previsti per target di categoria. Per i campi di target di categoria, un grafico a barre visualizza i valori previsti. I valori previsti vengono visualizzati come percentuale del campo obiettivo che viene previsto rientrare in ciascuna categoria. Per campi target di categoria con iterazioni di analisi della sensibilità, i risultati della categoria target prevista vengono visualizzati come grafico a barre con cluster che include i risultati di tutte le iterazioni. Il grafico viene suddiviso in cluster in base alla categoria o all'iterazione, a seconda dell'opzione selezionata nell'area **Raggruppamento per analisi sensibilità** nella vista Funzioni di densità della finestra di dialogo del nodo Valutazione simulazione. Fare clic con il tasto destro del mouse sul grafico per copiare, esportare o modificare il grafico. Selezionando **Modifica** il grafico viene aperto in una finestra mobile di editor con lavagna grafica.

Grafico Probabilità categoria previste per target di categoria. Per i campi target di categoria, un istogramma visualizza la distribuzione delle probabilità previste per ognuna delle categorie del target. Per i campi target di categoria con iterazioni di analisi della sensibilità, gli istogrammi vengono visualizzati in base alla categoria o all'iterazione, a seconda dell'opzione selezionata nell'area **Raggruppamento per analisi sensibilità** nella vista Funzioni di densità della finestra di dialogo del nodo Valutazione simulazione. Se gli istogrammi sono raggruppati in base alla categoria, un elenco a discesa contenente le etichette di iterazione consente di scegliere quale iterazione visualizzare. È inoltre possibile selezionare l'iterazione da visualizzare facendo clic con il tasto destro del mouse sul grafico e selezionando l'iterazione dal menu secondario **Iterazione**. Se gli istogrammi sono raggruppati in base all'iterazione, un elenco a discesa contenente i nomi categoria consente di scegliere quale categoria visualizzare. È inoltre possibile selezionare quale categoria visualizzare facendo clic con il tasto destro del mouse sul grafico e selezionando la categoria dal menu secondario **Categoria**.

Questo grafico è disponibile solo per un sottoinsieme di modelli e sul nugget del modello l'opzione per generare tutte le probabilità del gruppo deve essere selezionata. Ad esempio, sul nugget del modello Logistica, è necessario selezionare **Accoda tutte le probabilità**. I seguenti nugget del modello supportano questa opzione:

- Logistica, SVM, Bayes, Rete neurale e KNN
- Modelli di mining nel database DB2/ISW per regressione logistica, strutture ad albero delle decisioni e naïve Bayes

Per default, l'opzione per generare tutte le probabilità del gruppo non è selezionata su questi nugget del modello.

Grafici tornado. Il grafico tornado è un grafico a barre che visualizza la sensibilità del campo obiettivo rispetto ad ognuno degli input specificati. La sensibilità viene misurata dalla correlazione del target con ciascun input. Il titolo del grafico contiene il nome del campo obiettivo. Ogni barra nel grafico rappresenta la correlazione tra il campo obiettivo e un campo di input. Gli input simulati inclusi nel grafico sono gli input selezionati nel campo **Campi di input per tornado di correlazione** nella vista Campi della finestra di dialogo del nodo Valutazione simulazione. Ogni barra è etichettata con il valore di correlazione. Le barre sono ordinate in base al valore assoluto delle correlazioni, dal valore più grande al più piccolo. Se sono presenti iterazioni, viene generato un grafico separato per ogni iterazione. Ogni grafico ha un titolo secondario, che contiene il nome dell'iterazione.

Tabella di distribuzione. Questa tabella contiene il valore del campo obiettivo, sotto il quale rientra la percentuale di osservazioni specificata. La tabella contiene una riga per ogni valore di percentile specificato nella vista Output della finestra di dialogo del nodo Valutazione simulazione. I valori di percentile possono essere quartili, un numero diverso di percentili distanziati uniformemente, o percentili distanziati singolarmente. La tabella di distribuzione contiene una colonna per ciascuna iterazione.

Informazioni. Questa sezione fornisce un riepilogo complessivo dei campi e dei record utilizzati nella valutazione. Visualizza inoltre i campi di input ed i conteggi dei record, suddivisi per ogni iterazione.

Grafici: opzioni

Nella finestra di dialogo Opzioni grafico è possibile personalizzare la visualizzazione dei grafici attivati delle funzioni di densità di probabilità e di distribuzione cumulativa generate da una simulazione.

Vista. L'elenco a discesa **Vista** viene applicato solo al grafico della funzione di densità di probabilità. È possibile utilizzarlo per cambiare la vista grafico da una curva continua ad un istogramma. Questa funzione è disabilitata quando sullo stesso grafico vengono visualizzate più funzioni di densità (da più iterazioni). Quando sono presenti più funzioni di densità, possono essere visualizzate solo come curve continue.

Ordine. L'elenco a discesa **Ordine** viene applicato solo al grafico della funzione di distribuzione cumulativa. Specifica se la funzione di distribuzione cumulativa viene visualizzata come funzione crescente (default) o decrescente. Quando viene visualizzata come funzione decrescente, il valore della funzione in un determinato punto dell'asse orizzontale è la probabilità che il campo obiettivo sia alla destra di quel punto.

Posizioni dispositivo di scorrimento. Il campo di testo **Superiore** contiene la posizione corrente della linea di riferimento scorrevole destra. Il campo di testo **Inferiore** contiene la posizione corrente della linea di riferimento scorrevole sinistra. È possibile impostare esplicitamente le posizioni dei dispositivi di scorrimento immettendo i valori nei campi di testo **Superiore** e **Inferiore**. Il valore del campo di testo **Inferiore** deve essere necessariamente inferiore al valore contenuto nel campo di testo **Superiore**. È possibile rimuovere la linea di riferimento sinistra selezionando **-Infinito**, impostando di fatto la posizione su infinito negativo. Questa azione disabilita il campo di testo **Inferiore**. È possibile rimuovere la linea di riferimento destra selezionando **Infinito**, impostando di fatto la posizione su infinito. Questa azione disabilita il campo di testo **Superiore**. Non è possibile rimuovere entrambe le linee di riferimento; selezionando **-Infinito** si disabilita la casella di controllo **Infinito**, e viceversa.

Linee di riferimento. È possibile aggiungere diverse linee di riferimento verticali fisse alle funzioni di densità di probabilità e alle funzioni di distribuzione cumulativa.

- **Media.** È possibile aggiungere una linea di riferimento sul valore medio del campo obiettivo.
- **Mediana.** È possibile aggiungere una linea di riferimento sul valore mediano del campo obiettivo.
- **Deviazioni standard.** È possibile aggiungere linee di riferimento sul più e meno un numero specificato di deviazioni standard dalla media del campo obiettivo. È possibile immettere il numero di deviazioni standard da utilizzare nel campo di testo adiacente. Il numero minimo di deviazioni standard è 1 e il numero massimo è 10. Il numero di deviazioni standard viene inizialmente impostato su 3.
- **Percentili.** È possibile aggiungere linee di riferimento su uno o due valori di percentile della distribuzione del campo obiettivo immettendo i valori nei campi di testo **Inferiore** e **Superiore**. Ad esempio, il valore 95 nel campo di testo **Inferiore** rappresenta il 95° percentile, che è il valore sotto il quale rientra il 95% delle osservazioni. Allo stesso modo, il valore 5 nel campo di testo **Inferiore** rappresenta il 5° percentile, che è il valore sotto il quale rientra il 5% delle osservazioni. Per il campo di testo **Inferiore**, il valore di percentile minimo è 0 e il massimo è 49. Per il campo di testo **Superiore**, il valore di percentile minimo è 50 e il massimo è 100.
- **Posizioni personalizzate.** È possibile aggiungere linee di riferimento sui valori specificati lungo l'asse orizzontale. È possibile rimuovere le linee di riferimento personalizzate eliminando la voce dalla griglia.

Quando si fa clic su **OK**, i dispositivi di scorrimento, le etichette sopra i dispositivi di scorrimento, le linee di riferimento e la tabella sotto il grafico vengono aggiornati per riflettere le opzioni selezionate nella finestra di dialogo Opzioni grafico. Fare clic su **Annulla** per chiudere la finestra di dialogo senza apportare modifiche. Le linee di riferimento possono essere rimosse deselegzionando la scelta associata nella finestra di dialogo Opzioni grafico e facendo clic su **OK**.

Nota: Quando più funzioni di densità o di distribuzione vengono visualizzate su un singolo grafico (a causa dei risultati delle iterazioni di analisi della sensibilità), le linee di riferimento (diverse dalle linee personalizzate) vengono applicate separatamente a ciascuna funzione. Vengono visualizzate solo le linee

di riferimento della prima iterazione. Le etichette delle linee di riferimento includono l'etichetta dell'iterazione. L'etichetta dell'iterazione viene ricavata dall'upstream, in genere da un nodo Genera simulazione. Se non sono disponibili etichette dell'iterazione, viene utilizzato il valore dell'iterazione. Le opzioni **Media**, **Mediana**, **Deviazioni standard** e **Percentili** sono disabilitate per le funzioni di distribuzione cumulativa con più iterazioni.

IBM SPSS Statistics Applicazioni di supporto

Se sul computer è installata una versione compatibile di IBM SPSS Statistics con relativa licenza, è possibile configurare IBM SPSS Modeler in modo da elaborare i dati con la funzionalità IBM SPSS Statistics utilizzando i nodi Trasformazioni Statistics, Modelli Statistics, Output Statistics o Esporta Statistics.

Per ulteriori informazioni relative alla compatibilità del prodotto con la versione corrente di IBM SPSS Modeler, consultare il sito di supporto all'indirizzo <http://www.ibm.com/support>.

Per configurare IBM SPSS Modeler in modo che possa funzionare con IBM SPSS Statistics e altre applicazioni, scegliere:

Strumenti > Opzioni > Applicazione di supporto

IBM SPSS Statistics Interactive. Immettere il percorso completo e il nome del comando (per esempio, *C:\Programmi\IBM\SPSS\Statistics\<nn>\stats.exe*) da utilizzare per l'avvio di IBM SPSS Statistics direttamente in un file di dati prodotto dal nodo Esporta Statistics. Per ulteriori informazioni, consultare l'argomento "Nodo Esporta Statistics" a pagina 349.

Connessione. Se IBM SPSS Statistics Server e IBM SPSS Modeler Server sono installati nello stesso server, è possibile attivare una connessione tra le due applicazioni, in modo da incrementare il livello di efficienza mantenendo i dati sul server durante l'analisi. Selezionare **Server** per attivare l'opzione **Porta** riportata di seguito. L'impostazione di default è **Locale**.

Porta. Specificare la porta del server per IBM SPSS Statistics Server.

IBM SPSS Statistics Location Utility. Per consentire a IBM SPSS Modeler di utilizzare i nodi Trasformazioni Statistics, Modelli Statistics e Output Statistics, è necessario che sul computer dove viene eseguito il flusso sia installata una copia di IBM SPSS Statistics con relativa licenza.

- Se IBM SPSS Modeler è in esecuzione in modalità locale (autonoma), la copia con licenza di IBM SPSS Statistics deve essere presente sul computer locale. Fare clic su questo pulsante per specificare il percorso dell'installazione di IBM SPSS Statistics locale che si desidera utilizzare per la licenza.
- Inoltre, se in esecuzione in modalità distribuita su un IBM SPSS Modeler Server remoto, è necessario anche eseguire un programma di utilità sull'host IBM SPSS Modeler Server per creare il file *statistics.ini*, che indica a IBM SPSS Statistics il percorso di installazione di IBM SPSS Modeler Server. Per eseguire questa operazione, passare dal prompt dei comandi alla directory IBM SPSS Modeler Server *bin* e, per Windows, eseguire:

```
statisticsutility -location=<IBM SPSS Statistics_installation_path>/
```

In alternativa, per UNIX eseguire:

```
./statisticsutility -location=<IBM SPSS Statistics_installation_path>/bin
```

Se non si dispone di una copia con licenza di IBM SPSS Statistics sulla macchina locale, è possibile eseguire il nodo File Statistics su un server IBM SPSS Statistics, ma se vengono eseguiti tentativi di esecuzione di altri nodi IBM SPSS Statistics verrà visualizzato un messaggio di errore.

Commenti

In caso di problemi nell'esecuzione dei nodi Proc IBM SPSS Statistics, tenere presente quanto segue:

- Se i nomi dei campi utilizzati in IBM SPSS Modeler superano gli otto caratteri (per le versioni precedenti a IBM SPSS Statistics 12.0), superano i 64 caratteri (per IBM SPSS Statistics 12.0 e versioni successive) oppure contengono caratteri non validi, è necessario rinominarli o troncarli prima che possano essere letti in IBM SPSS Statistics. Per ulteriori informazioni, consultare l'argomento "Ridenominazione o filtro dei campi per IBM SPSS Statistics" a pagina 350.
- Se IBM SPSS Statistics è stato installato dopo IBM SPSS Modeler, potrebbe essere necessario specificare il percorso di IBM SPSS Statistics, come illustrato in precedenza.

Capitolo 7. Nodi di esportazione

Panoramica sui nodi di esportazione

I nodi di esportazione consentono di esportare i dati in vari formati per l'interfacciamento con gli altri strumenti software in uso.

Sono disponibili i seguenti nodi di esportazione:



Il nodo di esportazione del database scrive dati in una sorgente dati relazionale compatibile con ODBC. Per scrivere in una sorgente dati ODBC, è necessario utilizzare una sorgente dati esistente e disporre dell'autorizzazione in scrittura per tale sorgente.



Il nodo di esportazione File flat restituisce dati in un file di testo delimitato. È utile per esportare i dati che possono essere letti da altri software di analisi o fogli di calcolo.



Il nodo Esporta Statistics restituisce i dati in formato IBM SPSS Statistics *.sav* o *.zsav*. I file *.sav* o *.zsav* possono essere letti da IBM SPSS Statistics Base e da altri prodotti. Questo formato viene inoltre utilizzato per i file cache di IBM SPSS Modeler.



Il nodo Esportazione di IBM SPSS Data Collection esegue l'output di dati nel formato utilizzato dal software di ricerche di mercato IBM SPSS Data Collection. Per utilizzare questo nodo, è necessario che sia installata IBM SPSS Data Collection Data Library.



Il nodo di esportazione IBM Cognos BI esporta i dati in un formato leggibile dai database Cognos BI.



Il nodo di esportazione IBM Cognos TM1 esporta i dati in un formato che può essere letto dai database Cognos TM1.



Il nodo Esporta SAS restituisce nel formato SAS i dati che devono essere letti in SAS o in un pacchetto software compatibile con SAS. Sono disponibili tre formati di file SAS: SAS per Windows/OS2, SAS per UNIX o SAS Versione 7/8.



Il nodo Esportazione da Excel esegue l'output di dati in formato Microsoft Excel (*.xls*). Se lo si desidera, è possibile scegliere di avviare Excel automaticamente e aprire il file esportato quando si esegue il nodo.



Il nodo Esporta XML restituisce i dati in un file in formato XML. Se lo si desidera, è possibile creare un nodo origine XML per leggere nuovamente i dati esportati nel flusso.

Nodo di esportazione del database

È possibile usare i nodi Database per scrivere dati su sorgenti di dati relazionali compatibili con ODBC, come illustrato nella descrizione del nodo origine del database. Per ulteriori informazioni, consultare l'argomento "Nodo origine del database" a pagina 13.

Per scrivere dati in un database, utilizzare i passaggi seguenti:

1. Installare un driver ODBC e configurare una sorgente dati per il database che si desidera utilizzare.
2. Nella scheda Esporta del nodo Database, specificare la sorgente dati e la tabella in cui scrivere i dati. È possibile creare una nuova tabella o inserire i dati in una tabella esistente.
3. Se necessario, impostare altre opzioni.

Questi passaggi verranno descritti in dettaglio negli argomenti seguenti.

Scheda di esportazione del nodo Database

Origine dati. Mostra la sorgente dati selezionata. Immettere il nome o selezionarlo dall'elenco a discesa. Se nell'elenco non compare il database desiderato, selezionare **Aggiungi nuova connessione database** e individuare il database dalla finestra di dialogo Connessioni database. Per ulteriori informazioni, consultare l'argomento "Aggiunta di una connessione a un database" a pagina 15.

Nome tabella. Immettere il nome della tabella alla quale si desidera inviare i dati. Se si seleziona l'opzione **Inserisci in tabella**, è possibile selezionare una tabella esistente del database facendo clic sul pulsante **Seleziona**.

Crea tabella. Selezionare questa opzione per creare una nuova tabella di database o per sovrascriverne una esistente.

Inserisci in tabella. Selezionare questa opzione per inserire i dati come nuove righe in una tabella di database esistente.

Unisci tabella. (Dove disponibile) Selezionare questa opzione per aggiornare le colonne del database selezionate con i valori dei campi dati di origine corrispondenti. La selezione di questa opzione abilita il pulsante **Unione**, che visualizza una finestra di dialogo da cui è possibile mappare i campi dati di origine alle colonne del database.

Elimina tabella esistente. Selezionare questa opzione per eliminare qualsiasi tabella esistente con lo stesso nome della nuova tabella che si desidera creare.

Elimina righe esistenti. Selezionare questa opzione per eliminare dalla tabella le righe esistenti prima dell'esportazione quando si inseriscono dati in una tabella.

Nota: se è selezionata una delle due opzioni sopra riportate, durante l'esecuzione del nodo verrà visualizzato un messaggio **Avviso di sovrascrittura**. Per eliminare gli avvisi, deselezionare **Avvisa quando un nodo sovrascrive una tabella di database** nella scheda Notifiche della finestra di dialogo Opzioni utente.

Dimensioni di default stringhe. I campi contrassegnati come senza tipo in un nodo Tipo a monte vengono scritti nel database come campi stringa. Specificare la dimensione delle stringhe da utilizzare per i campi senza tipo.

Fare clic su **Schema** per visualizzare una finestra di dialogo in cui è possibile configurare diverse opzioni di esportazione (per i database che supportano tale funzione), impostare i tipi di dati SQL per i campi e specificare la chiave primaria ai fini dell'indicizzazione del database. Per ulteriori informazioni, consultare l'argomento "Esportazione Database - Opzioni Schema" a pagina 318.

Fare clic su **Indici** per impostare le opzioni di indicizzazione della tabella esportata al fine di migliorare le prestazioni del database. Per ulteriori informazioni, consultare l'argomento "Esportazione database - Opzioni Indice" a pagina 321.

Fare clic su **Opzioni avanzate** per specificare le opzioni di caricamento di massa e di commit nel database. Per ulteriori informazioni, consultare l'argomento "Esportazione database - Opzioni avanzate" a pagina 323.

Nomi di tabelle e colonne fra virgolette. Selezionare le opzioni da utilizzare al momento dell'invio di un'istruzione CREATE TABLE al database. Le tabelle e le colonne con spazi o caratteri non standard devono essere racchiuse tra virgolette.

- **Se necessario.** Selezionare questa opzione per consentire a IBM SPSS Modeler di determinare automaticamente l'eventuale esigenza di virgolette per ogni singolo caso.
- **Sempre.** Selezionare questa opzione per racchiudere sempre i nomi di tabella e di colonna tra virgolette.
- **Mai.** Selezionare questa opzione per disattivare l'utilizzo delle virgolette.

Genera un nodo di importazione per questi dati. Selezionare questa opzione per generare un nodo origine del database per i dati esportati nella sorgente dati e nella tabella specificate. Dopo l'esecuzione, questo nodo verrà aggiunto all'area del flusso.

Opzioni di unione esportazione database

Questa finestra di dialogo consente di mappare i campi dai dati di origine alle colonne della tabella di database di destinazione. Se un campo dati di origine viene mappato a una colonna del database, quando si esegue il flusso il valore della colonna viene sostituito dal valore dei dati di origine. I campi di input non mappati restano invariati nel database.

Mappa campi. Qui è possibile specificare la mappatura tra i campi dati di origine e le colonne del database. I campi dati di origine con lo stesso nome delle colonne del database vengono mappati automaticamente.

- **Mappa.** Mappa un campo dati di origine selezionato nell'elenco dei campi a sinistra del pulsante a una colonna del database selezionata nell'elenco a destra. È possibile mappare più campi contemporaneamente, ma il numero delle voci selezionate in entrambi gli elenchi deve essere lo stesso.
- **Annulla mappatura.** Elimina la mappatura per una o più colonne del database selezionate. Il pulsante viene attivato quando si seleziona un campo o una colonna di database nella tabella a destra della finestra di dialogo.
- **Aggiungi.** Aggiunge uno o più campi dati di origine selezionati nell'elenco dei campi a sinistra del pulsante all'elenco di destra dei campi da mappare. Il pulsante viene attivato quando si seleziona un campo nell'elenco a sinistra e non è presente alcun campo con lo stesso nome nell'elenco a destra. Se si fa clic sul pulsante, il campo selezionato viene mappato su una nuova colonna di database con lo stesso nome. Il termine <NEW> viene visualizzato dopo il nome della colonna di database a indicare che si tratta di un nuovo campo.

Unisci righe. Utilizza un campo chiave, per esempio *ID transazione*, per unire record con lo stesso valore del campo chiave. Questa opzione è equivalente a una "unione di uguaglianza" (equi-join) di database. I valori chiave devono essere quelli delle chiavi primarie, ovvero devono essere univoci e non possono contenere valori null.

- **Chiavi possibili.** Elenca tutti i campi trovati in tutte le origini dati di input. Selezionare uno o più campi dall'elenco e utilizzare il pulsante con la freccia per aggiungerli come campi chiave da utilizzare

per l'unione dei record. Tutti i campi della mappa a cui corrisponde una colonna del database mappata possono essere utilizzati come chiave, tranne quelli aggiunti come nuove colonne del database (contraddistinti da <NEW> dopo il nome).

- **Chiavi per l'unione.** Elenca tutti i campi utilizzati per unire i record da tutte le origini dati di input in base ai valori dei campi chiave. Per rimuovere una chiave dall'elenco, selezionarla e utilizzare il pulsante con la freccia per includerla nuovamente nell'elenco Chiavi possibili. L'opzione seguente è attivata solo quando vengono selezionati più campi.
- **Includi solo record esistenti nel database.** Esegue un'unione parziale; se il record è presente sia nel database che nel flusso, i campi mappati vengono aggiornati.
- **Aggiungi record al database.** Esegue un'unione esterna; tutti i record del flusso vengono uniti (se nel database è presente lo stesso record) o aggiunti (se il record non esiste ancora nel database).

Per mappare un campo dati di origine a una nuova colonna del database

1. Fare clic sul nome del campo di input nell'elenco a sinistra, sotto **Mappa campi**.
2. Fare clic sul pulsante **Aggiungi** per completare la mappatura.

Per mappare un campo dati di origine a una colonna del database esistente

1. Fare clic sul nome del campo di input nell'elenco a sinistra, sotto **Mappa campi**.
2. Fare clic sul nome della colonna in **Colonna di database** sulla destra.
3. Fare clic sul pulsante **Mappa** per completare la mappatura.

Per rimuovere una mappatura

1. Nell'elenco a destra, sotto Campo, fare clic sul nome del campo di cui si desidera rimuovere la mappatura.
2. Fare clic sul pulsante **Annulla mappatura**.

Per deselezionare un campo in uno degli elenchi

Tenere premuto il tasto CTRL e fare clic sul nome del campo.

Esportazione Database - Opzioni Schema

Nella finestra di dialogo Schema dell'esportazione del database, è possibile impostare le opzioni per l'esportazione del database (per i database che supportano tali opzioni), impostare i tipi di dati SQL per i propri campi, specificare quali campi rappresentano chiavi primarie e personalizzare l'istruzione CREATE TABLE generata durante l'esportazione.

La finestra di dialogo è composta da varie parti:

- La sezione superiore (se visualizzata) include le opzioni di esportazione in un database che supporta tali opzioni. Questa sezione non è visualizzata se non si è collegati a un database di questo tipo.
- Il campo testo al centro visualizza il modello utilizzato per generare il comando CREATE TABLE, che per default segue il formato:
`CREATE TABLE <nome-tabella> <(colonne tabella)>`
- La tabella nella parte inferiore consente di specificare il tipo di dati SQL per ciascun campo e di indicare quali campi sono chiavi primarie, come illustrato di seguito. La finestra di dialogo genera automaticamente i valori dei parametri <nome-tabella> e <(colonne tabella)> in base alle specifiche contenute nella tabella.

Impostazione delle opzioni di esportazione in database

Se la sezione è visualizzata, è possibile specificare diverse impostazioni di esportazione nel database. Questa funzione è supportata dai tipi di database elencati di seguito.

- IBM InfoSphere Warehouse eseguito su DB2 9.1 o successiva. Per ulteriori informazioni, consultare l'argomento "Opzioni per IBM DB2 InfoSphere Warehouse".
- SQL Server 2008 o le successive versioni Enterprise e Developer. Per ulteriori informazioni, consultare l'argomento "Opzioni per SQL Server" a pagina 320.
- Oracle 10g e 11gR1 o le successive versioni Enterprise o Personal. Per ulteriori informazioni, consultare l'argomento "Opzioni per Oracle" a pagina 320.

Personalizzazione delle istruzioni CREATE TABLE

Utilizzando la parte del campo di testo di questa finestra di dialogo, è possibile aggiungere ulteriori opzioni specifiche del database all'istruzione CREATE TABLE.

1. Selezionare la casella di controllo **Personalizza comando CREATE TABLE** per attivare la finestra di testo.
2. Aggiungere all'istruzione le opzioni specifiche del database desiderate. Assicurarsi di mantenere i parametri di testo <nome-tabella> e (<colonne-tabella>) perché verranno sostituiti dal nome effettivo della tabella e dalle definizioni delle colonne in IBM SPSS Modeler.

Impostazione dei tipi di dati SQL

Per default, IBM SPSS Modeler abilita il server di database all'assegnazione automatica dei tipi di dati SQL. Per ignorare il tipo automatico per un campo, trovare la riga corrispondente al campo e selezionare il tipo desiderato dall'elenco a discesa nella colonna *Tipo* della tabella dello schema. Per selezionare più righe è possibile utilizzare la combinazione Maiusc-clic.

Se ai tipi è associato un argomento di lunghezza, di precisione o di scala (BINARY, VARBINARY, CHAR, VARCHAR, NUMERIC e NUMBER), è consigliabile specificare una lunghezza anziché consentire al server di database di assegnare una lunghezza automatica. Se lo si desidera, ad esempio, è possibile garantire la sovrascrittura del tipo di archiviazione in IBM SPSS Modeler specificando un valore sensibile per la lunghezza, come VARCHAR(25). Per ignorare l'assegnazione automatica, selezionare **Specifica** dall'elenco a discesa Tipo e sostituire la definizione del tipo con l'istruzione di definizione del tipo SQL desiderato.

Il modo più semplice per modificare l'argomento per un tipo consiste nel selezionare innanzitutto il tipo maggiormente corrispondente alla definizione del tipo desiderato e quindi **Specifica** per modificare tale definizione. Per esempio, per impostare il tipo di dati SQL su VARCHAR(25), impostare innanzitutto il tipo su **VARCHAR(lunghezza)** nell'elenco a discesa Tipo, quindi selezionare **Specifica** e sostituire la lunghezza del testo con il valore 25.

Chiavi primarie

Se una o più colonne della tabella esportata devono avere un valore o una combinazione di valori univoci per ogni riga, è possibile indicarlo selezionando la casella di controllo **Chiave primaria** per ogni campo valido. La maggior parte dei database non consente di apportare modifiche alla tabella che possano invalidare il vincolo della chiave primaria e crea automaticamente un indice sulla chiave primaria per favorire il rispetto di questa restrizione. Se lo si desidera, è possibile creare indici per altri campi nella finestra di dialogo Indici. Per ulteriori informazioni, consultare l'argomento "Esportazione database - Opzioni Indice" a pagina 321.

Opzioni per IBM DB2 InfoSphere Warehouse

Spazio di tabella. Lo spazio di tabella da utilizzare per l'esportazione. Gli amministratori di database possono creare o configurare spazi di tabella partizionati. Si consiglia di selezionare uno di questi spazi di tabella (anziché quello di default) da utilizzare per le esportazioni in database.

Partiziona i dati per campo. Specifica il campo di input da utilizzare per il partizionamento.

Utilizza compressione. Se questa opzione è selezionata, vengono create le tabelle per l'esportazione con la compressione (ad esempio, l'equivalente di CREATE TABLE MYTABLE(...) COMPRESS YES; in SQL).

Opzioni per SQL Server

Utilizza compressione. Se è selezionata, questa opzione utilizza la compressione nella creazione delle tabelle per l'esportazione.

Compressione per. Scegliere il livello di compressione.

- **Riga.** Abilita la compressione a livello di riga (ad esempio, l'equivalente di CREATE TABLE MYTABLE(...) WITH (DATA_COMPRESSION = ROW); in SQL).
- **Pagina.** Abilita la compressione a livello di pagina (ad esempio, CREATE TABLE MYTABLE(...) WITH (DATA_COMPRESSION = PAGE); in SQL).

Opzioni per Oracle

Impostazioni di Oracle 10g

Utilizza compressione. Se è selezionata, questa opzione utilizza la compressione nella creazione delle tabelle per l'esportazione. Per questa versione del database, è disponibile solo la compressione di base (ad esempio, CREATE TABLE MYTABLE(...) COMPRESS; in SQL).

Impostazioni di Oracle 11gR1

Utilizza compressione. Se è selezionata, questa opzione utilizza la compressione nella creazione delle tabelle per l'esportazione.

Compressione per. Scegliere il livello di compressione.

- **Default.** Abilita la compressione di default (ad esempio, CREATE TABLE MYTABLE(...) COMPRESS; in SQL). In questo caso, svolge la stessa funzione dell'opzione **Operazioni caricamento diretto**.
- **Operazioni caricamento diretto.** Abilita la compressione solo per le operazioni di inserimento di massa (percorso diretto) (ad esempio, CREATE TABLE MYTABLE(...) COMPRESS FOR DIRECT_LOAD OPERATIONS; in SQL).
- **Tutte le operazioni.** Abilita la compressione per tutte le operazioni (ad esempio, CREATE TABLE MYTABLE(...) COMPRESS FOR ALL OPERATIONS; in SQL).

Impostazioni di Oracle 11gR2 - Opzione di base

Utilizza compressione. Se è selezionata, questa opzione utilizza la compressione nella creazione delle tabelle per l'esportazione.

Compressione per. Scegliere il livello di compressione.

- **Default.** Abilita la compressione di default (ad esempio, CREATE TABLE MYTABLE(...) COMPRESS; in SQL). In questo caso, svolge la stessa funzione dell'opzione **Di base**.
- **Di base.** Abilita la compressione di base (ad esempio, CREATE TABLE MYTABLE(...) COMPRESS BASIC; in SQL).

Impostazioni di Oracle 11gR2 - Opzione avanzata

Utilizza compressione. Se è selezionata, questa opzione utilizza la compressione nella creazione delle tabelle per l'esportazione.

Compressione per. Scegliere il livello di compressione.

- **Default.** Abilita la compressione di default (ad esempio, CREATE TABLE MYTABLE(...) COMPRESS; in SQL). In questo caso, svolge la stessa funzione dell'opzione **Di base**.

- **Di base.** Abilita la compressione di base (ad esempio, CREATE TABLE MYTABLE(...) COMPRESS BASIC; in SQL).
- **OLTP.** Abilita la compressione OLTP (ad esempio, CREATE TABLE MYTABLE(...)COMPRESS FOR OLTP; in SQL).
- **Min/Max query.** (Solo server Exadata) Abilita la compressione colonnare ibrida per la query (ad esempio, CREATE TABLE MYTABLE(...)COMPRESS FOR QUERY LOW; o CREATE TABLE MYTABLE(...)COMPRESS FOR QUERY HIGH; in SQL). La compressione per la query è utile negli ambienti data warehousing; HIGH fornisce un rapporto di compressione maggiore rispetto a LOW.
- **Min/Max archivio.** (Solo server Exadata) Abilita la compressione colonnare ibrida per l'archivio (ad esempio, CREATE TABLE MYTABLE(...)COMPRESS FOR ARCHIVE LOW; o CREATE TABLE MYTABLE(...)COMPRESS FOR ARCHIVE HIGH; in SQL). La compressione per l'archivio è utile per la compressione dei dati che verranno archiviati per lunghi periodi di tempo; HIGH fornisce un rapporto di compressione maggiore rispetto a LOW.

Esportazione database - Opzioni Indice

La finestra di dialogo Indici consente di creare indici sulle tabelle di database esportate da IBM SPSS Modeler. È possibile specificare gli insiemi di campi da includere e personalizzare il comando CREATE INDEX in funzione delle proprie esigenze.

La finestra di dialogo è composta da due parti:

- Il campo testo nella parte superiore visualizza un modello che può essere utilizzato per generare uno o più comandi CREATE INDEX, che per default segue il formato:
CREATE INDEX <nome-indice> ON <nome-tabella>
- La tabella nella parte inferiore della finestra di dialogo consente di aggiungere le specifiche per ciascun indice che si desidera creare. Per ogni indice vengono specificati il nome e i campi o le colonne da includere. La finestra di dialogo genera automaticamente i valori dei parametri <nome-indice> e <nome-tabella>.

Per esempio, la stringa SQL generata per un singolo indice sui campi *empid* e *deptid* potrebbe avere il seguente aspetto:

```
CREATE INDEX MYTABLE_IDX1 ON MYTABLE(EMPID,DEPTID)
```

È possibile aggiungere più righe per creare indici multipli. Per ogni riga viene generato un comando CREATE INDEX separato.

Personalizzazione del comando CREATE INDEX

Se lo si desidera, è possibile personalizzare il comando CREATE INDEX per tutti gli indici o solo per un indice specifico. Questo garantisce la flessibilità necessaria per adeguarsi ai requisiti o alle opzioni specifiche di un database e per applicare le personalizzazioni a tutti gli indici o solo a indici specifici, in funzione delle proprie esigenze.

- Selezionare **Personalizza comando CREATE INDEX** nella parte superiore della finestra di dialogo per modificare il modello utilizzato per tutti gli indici aggiunti in seguito. Si noti che le modifiche non vengono applicate automaticamente agli indici che sono già stati aggiunti alla tabella.
- Selezionare una o più righe della tabella e fare clic su **Aggiorna indici selezionati** nella parte superiore della finestra di dialogo per applicare le personalizzazioni correnti a tutte le righe selezionate.
- Selezionare la casella di controllo **Personalizza** in ogni riga per modificare il modello del comando solo per quell'indice.

Si noti che i valori dei parametri <nome-indice> e <nome-tabella> sono generati automaticamente dalla finestra di dialogo a seconda delle specifiche della tabella e non possono essere modificati direttamente.

BITMAP KEYWORD. Se si utilizza un database di Oracle, è possibile personalizzare il modello per creare un indice bitmap anziché un indice standard, nel modo seguente:

```
CREATE BITMAP INDEX <nome-indice> ON <nome-tabella>
```

Gli indici bitmap possono essere utili per indicizzare le colonne con un numero limitato di valori distinti. La stringa SQL ottenuta potrebbe avere il seguente aspetto:

```
CREATE BITMAP INDEX MYTABLE_IDX1 ON MYTABLE(COLOR)
```

Parola chiave UNIQUE. La maggior parte dei database supporta la parola chiave UNIQUE nel comando CREATE INDEX. Questo applica un vincolo di univocità simile a quello della chiave primaria nella tabella sottostante.

```
CREATE UNIQUE INDEX <nome-indice> ON <nome-tabella>
```

Si noti che, per i campi effettivamente designati come chiavi primarie, questa specifica non è necessaria. La maggior parte dei database crea automaticamente un indice per tutti i campi designati come chiavi primarie nel comando CREATE TABLE, per cui non è necessario creare esplicitamente degli indici in questi campi. Per ulteriori informazioni, consultare l'argomento "Esportazione Database - Opzioni Schema" a pagina 318.

Parola chiave FILLFACTOR. Alcuni parametri fisici dell'indice possono essere ulteriormente raffinati. Ad esempio, SQL Server consente all'utente di raggiungere un compromesso tra le dimensioni dell'indice (dopo la creazione iniziale) ed i costi di manutenzione delle future modifiche apportate alla tabella.

```
CREATE INDEX MYTABLE_IDX1 ON MYTABLE(EMPID,DEPTID) WITH FILLFACTOR=20
```

Altre osservazioni

- Se esiste già un indice con il nome specificato, la creazione dell'indice non andrà a buon fine. Gli eventuali errori vengono inizialmente considerati avvisi, consentendo la creazione degli indici successivi e la nuova segnalazione dell'errore nel registro dei messaggi dopo che il database ha tentato di creare tutti gli indici.
- Per ottimizzare le prestazioni, gli indici devono essere creati dopo avere caricato i dati nella tabella. Gli indici devono contenere almeno una colonna.
- Prima di eseguire il nodo è possibile visualizzare un'anteprima della stringa SQL generata nel registro dei messaggi.
- Per le tabelle temporanee che vengono scritte nel database (quando cioè viene attivata la cache dei nodi) non sono disponibili le opzioni che consentono di indicare le chiavi primarie e gli indici. Tuttavia, se necessario, il sistema può creare degli indici nella tabella temporanea, a seconda di come vengono utilizzati i dati nei nodi downstream. Per esempio, se ai dati memorizzati nella cache si aggiunge in seguito una colonna *DEPT*, sarebbe sensato indicizzare la tabella memorizzata nella cache in questa colonna.

Indici e ottimizzazione delle query

In alcuni sistemi di gestione di database, dopo che una tabella di database è stata creata, caricata e indicizzata è necessaria un'ulteriore operazione per consentire all'ottimizzatore di utilizzare gli indici per accelerare l'esecuzione delle query nella nuova tabella. Per esempio, in Oracle, l'ottimizzatore di query basato sui costi richiede che la tabella venga prima analizzata per poterne utilizzare gli indici per l'ottimizzazione delle query. Il file delle proprietà ODBC interno per Oracle (non visibile all'utente) contiene un'opzione perché questo accada, nel modo seguente:

```
# Defines SQL to be executed after a table and any associated indexes  
# have been created and populated  
table_analysis_sql, 'ANALYZE TABLE <table-name> COMPUTE STATISTICS'
```


Questa operazione viene eseguita tutte le volte che viene creata una tabella in Oracle (indipendentemente dal fatto che vengano definite delle chiavi primarie o degli indici). Se necessario, il file delle proprietà ODBC per database aggiuntivi può essere personalizzato in modo analogo. Contattare l'assistenza.

Esportazione database - Opzioni avanzate

Facendo clic sul pulsante Avanzate nella finestra di dialogo Nodo di esportazione del database, viene visualizzata una nuova finestra di dialogo in cui è possibile specificare i dettagli tecnici per l'esportazione dei risultati in un database.

Utilizza commit di batch. Selezionare questa opzione per disattivare i commit riga per riga nel database.

Dimensione batch. Specificare il numero di record da inviare al database prima del commit nella memoria. La riduzione di questo numero garantisce una maggiore integrità dei dati, rallentando però la velocità di trasferimento. Può rivelarsi utile regolare questo numero in modo da ottenere prestazioni ottimali con il database in uso.

Opzioni InfoSphere Warehouse. Visualizzata solo se si è collegati a un database InfoSphere Warehouse (IBM DB2 9.7 o versioni successive). **Non registrare gli aggiornamenti** consente di evitare di registrare gli eventi quando si creano le tabelle e si inseriscono i dati.

Utilizza caricamento di massa. Selezionare un metodo per eseguire il caricamento di massa dei dati nel database direttamente da IBM SPSS Modeler. Può essere necessario fare qualche esperimento per determinare le opzioni di caricamento di massa più appropriate per una determinata situazione.

- **Via ODBC.** Selezionare questa opzione per utilizzare l'API ODBC in modo da eseguire inserimenti di più righe con una maggiore efficienza rispetto alla normale esportazione nel database. Nelle opzioni riportate di seguito scegliere l'associazione per riga o per colonna.
- **Via programma di caricamento esterno.** Selezionare questa opzione per utilizzare un programma di caricamento di massa personalizzato, specifico per il database in uso. Selezionando questa opzione verranno attivate le diverse opzioni riportate di seguito.

Opzioni ODBC avanzate. Queste opzioni sono disponibili soltanto in caso di selezione di **Via ODBC**. Si noti che questa funzionalità potrebbe non essere supportata da tutti i driver ODBC.

- **Per riga.** Selezionare l'associazione per riga per utilizzare la chiamata a `SQLBulkOperations` per il caricamento dei dati nel database. L'associazione per riga garantisce in genere una velocità più elevata rispetto all'utilizzo di inserimenti con parametri, con cui i dati vengono inseriti record per record.
- **Per colonna.** Selezionare questa opzione per utilizzare l'associazione per colonna per il caricamento dei dati nel database. L'associazione per colonna garantisce un incremento delle prestazioni associando ogni colonna del database (in un'istruzione `INSERT` con parametri) a un array di valori N . Se l'istruzione `INSERT` viene eseguita una sola volta, nel database vengono inserite N righe. Questo metodo può determinare un notevole incremento delle prestazioni.

Opzioni programma di caricamento esterno. Quando si specifica **Via programma di caricamento esterno**, vengono visualizzate diverse opzioni che consentono di esportare l'insieme di dati in un file e di specificare ed eseguire un programma di caricamento personalizzato per il caricamento dei dati da tale file al database. IBM SPSS Modeler può interfacciarsi con i programmi di caricamento esterni di molti dei più diffusi sistemi di database. Il software è stato corredato di numerosi script che sono disponibili insieme alla documentazione tecnica nella sottodirectory *scripts*. Notare che per utilizzare questa funzione, è necessario che sia installato Python 2.7 sulla stessa macchina su cui è installato IBM SPSS Modeler o IBM SPSS Modeler Server e che il parametro `python_exe_path` sia impostato nel file *options.cfg*. Per ulteriori informazioni, consultare l'argomento "Programmi di caricamento di massa" a pagina 324.

- **Utilizza delimitatore.** Specificare il carattere di delimitazione da utilizzare nel file esportato. Selezionare **Tabulatore** per eseguire la delimitazione tramite tabulazione e **Spazio** per delimitare tramite spazi. Selezionare **Altro** per specificare un altro carattere, per esempio una virgola (,).

- **Specifica file di dati.** Selezionare questa opzione per immettere il percorso da utilizzare per il file di dati scritto durante il caricamento di massa. Per default, viene creato un file temporaneo nella directory TEMP sul server.
- **Specifica programma di caricamento.** Selezionare questa opzione per specificare un programma per il caricamento di massa. Per default, viene cercato uno script Python da eseguire per un dato database nella sottodirectory *scripts* dell'installazione di IBM SPSS Modeler. Il software è stato corredato di numerosi script che sono disponibili insieme alla documentazione tecnica nella sottodirectory *scripts*.
- **Genera registro.** Selezionare questa opzione per generare un file di registro nella directory specificata. Il file di registro, contenente informazioni sugli errori, si rivela utile in caso di esito negativo dell'operazione di caricamento di massa.
- **Controlla dimensione tabella.** Selezionare questa opzione per eseguire il controllo della tabella e verificare così che l'incremento di dimensione della tabella corrisponda al numero di righe esportato da IBM SPSS Modeler.
- **Opzioni programma di caricamento esterno.** Specificare argomenti aggiuntivi per il programma di caricamento. Per gli argomenti contenenti spazi, utilizzare le virgolette.

Per includere le virgolette negli argomenti facoltativi, utilizzare una barra rovesciata come carattere di escape. Ad esempio, l'opzione specificata come `-comment "This is a \"comment\""` include l'indicatore `-comment` ed il commento stesso indicato come `This is a "comment"`.

È possibile includere una singola barra rovesciata utilizzando un'altra barra rovesciata come carattere di escape. Ad esempio, l'opzione specificata come `-specialdir "C:\\Test Scripts\\"` include l'indicatore `-specialdir` e la directory indicata come `C:\Test Scripts\`.

Programmi di caricamento di massa

Il nodo di esportazione del database prevede delle opzioni per il caricamento di massa nella finestra di dialogo Opzioni avanzate. I programmi di caricamento di massa possono essere utilizzati per caricare i dati in un database da un file di testo.

Con l'opzione **Utilizza caricamento di massa - Via programma di caricamento esterno**, IBM SPSS Modeler viene configurato per eseguire tre operazioni:

- creare tutte le tabelle di database necessarie;
- esportare i dati in un file di testo;
- richiamare un programma di caricamento di massa per caricare i dati del file nella tabella di database.

In genere il programma di caricamento di massa non è il programma di utilità di caricamento del database (per esempio il programma di utilità `sqlldr` di Oracle), ma un piccolo script o programma che elabora gli argomenti corretti, crea gli eventuali file ausiliari specifici del database (per esempio un file di controllo) e quindi richiama il programma di utilità di caricamento del database. Le informazioni contenute nelle sezioni seguenti illustrano come modificare un programma di caricamento di massa esistente.

In alternativa è possibile scrivere il proprio programma di caricamento di massa. Per ulteriori informazioni, consultare l'argomento "Sviluppo di programmi di caricamento di massa" a pagina 328.

Script per il caricamento di massa

IBM SPSS Modeler viene fornito con una serie di programmi di caricamento di massa, che sono implementati mediante gli script Python, per diversi database. Quando si esegue un flusso contenente un nodo di esportazione del database con l'opzione **Via programma di caricamento esterno** selezionata, IBM SPSS Modeler crea la tabella di database (se necessario) con ODBC, esporta i dati in un file temporaneo sull'host su cui è in esecuzione IBM SPSS Modeler Server, quindi richiama lo script di caricamento di massa. Questo script esegue a sua volta i programmi di utilità offerte dal fornitore DBMS per caricare i dati dai file temporanei al database.

Nota: l'installazione IBM SPSS Modeler non include un interprete di runtime Python, per cui è necessaria un'installazione di Python separata. Per ulteriori informazioni, consultare l'argomento "Esportazione database - Opzioni avanzate" a pagina 323.

Vengono forniti gli script (nella cartella `\scripts` della directory di installazione di IBM SPSS Modeler) per i database elencati nella tabella riportata di seguito.

Tabella 53. Script del programma di caricamento di massa forniti.

Database	Nome script	Ulteriori informazioni
IBM DB2	<code>db2_loader.py</code>	Per ulteriori informazioni, consultare l'argomento "Caricamento di massa dei dati nei database IBM DB2".
IBM Netezza	<code>netezza_loader.py</code>	Per ulteriori informazioni, consultare l'argomento "Caricamento di massa dei dati nei database IBM Netezza" a pagina 326.
Oracle	<code>oracle_loader.py</code>	Per ulteriori informazioni, consultare l'argomento "Caricamento di massa dei dati in database Oracle" a pagina 326.
SQL Server	<code>mssql_loader.py</code>	Per ulteriori informazioni, consultare l'argomento "Caricamento di massa dei dati nei database SQL Server" a pagina 327.
Teradata	<code>teradata_loader.py</code>	Per ulteriori informazioni, consultare l'argomento "Caricamento di massa dei dati in database Teradata" a pagina 328.

Caricamento di massa dei dati nei database IBM DB2

I seguenti punti potrebbero essere d'aiuto nella configurazione del caricamento di massa da IBM SPSS Modeler a un database IBM DB2 mediante l'opzione Programma di caricamento esterno nella finestra di dialogo Esportazione DB: Opzioni avanzate.

Assicurarsi che sia installato il programma di utilità CLP (command line processor, ovvero processore dalla riga di comando) DB2

Lo script `db2_loader.py` richiama il comando DB2 LOAD. Assicurarsi che il processore da riga di comando (`db2` in UNIX, `db2cmd` in Windows) sia installato sul server su cui deve essere eseguito `db2_loader.py` (solitamente, l'host su cui è in esecuzione IBM SPSS Modeler Server).

Controllare se l'alias del database locale è identico al nome effettivo del database

L'alias del database locale DB2 è il nome utilizzato dal software client DB2 per fare riferimento a un database in un'istanza DB2 locale o remota. Se l'alias del database locale è diverso dal nome del database remoto, fornire l'opzione del programma di caricamento esterno:

```
-alias <local_database_alias>
```

Ad esempio, il database remoto è denominato STARS sull'host GALAXY ma l'alias del database locale DB2 sull'host su cui è in esecuzione IBM SPSS Modeler Server è STARS_GALAXY. Utilizzare l'opzione del programma di caricamento esterno

```
-alias STARS_GALAXY
```

Codifica dei dati in caratteri non ASCII

Se si stanno caricando in massa dati che non sono in formato ASCII, accertarsi che la variabile codepage nella sezione di configurazione di `db2_loader.py` sia configurata correttamente nel sistema.

Stringhe vuote

Le stringhe vuote sono esportate nel database come valori NULL.

Caricamento di massa dei dati nei database IBM Netezza

I seguenti punti potrebbero essere d'aiuto nella configurazione del caricamento di massa da IBM SPSS Modeler a un database IBM Netezza mediante l'opzione Programma di caricamento esterno nella finestra di dialogo Esportazione DB: Opzioni avanzate.

Assicurarsi che il programma di utilità Netezza *nzload* sia installato

Lo script *netezza_loader.py* richiama il programma di utilità Netezza *nzload*. Assicurarsi che *nzload* sia installata e configurata correttamente sul server su cui deve essere eseguito *netezza_loader.py*.

Esportazione dei dati non ASCII

Se l'esportazione contiene dati non in formato ASCII, potrebbe essere necessario aggiungere `-encoding UTF8` al campo **Opzioni programma di caricamento esterno** nella finestra di dialogo Esportazione database - Opzioni avanzate. In questo modo i dati non ASCII vengono caricati correttamente.

Dati in formato Data, Ora e Timestamp

Nelle proprietà del flusso, impostare il formato della data su **GG-MM-AAAA** e il formato dell'ora su **HH:MM:SS**.

Stringhe vuote

Le stringhe vuote sono esportate nel database come valori NULL.

Ordine diverso di colonne nel flusso e nella tabella di destinazione quando si inseriscono i dati in una tabella esistente

Se l'ordine delle colonne nel flusso è diverso da quello nella tabella di destinazione, i valori dei dati verranno inseriti nelle colonne sbagliate. Utilizzare un nodo Ordina campi per assicurarsi che l'ordine delle colonne nel flusso corrisponda all'ordine nella tabella di destinazione. Per ulteriori informazioni, consultare l'argomento "Nodo Ordina campi" a pagina 163.

Monitoraggio dell'avanzamento di *nzload*

Quando IBM SPSS Modeler viene eseguito in modalità locale, aggiungere `-sts` al campo **Opzioni programma di caricamento esterno** nella finestra di dialogo Esportazione database - Opzioni avanzate per visualizzare i messaggi di stato ogni 10000 righe nella finestra di comando aperta del programma di utilità *nzload*.

Caricamento di massa dei dati in database Oracle

I seguenti punti potrebbero essere d'aiuto nella configurazione del caricamento di massa da IBM SPSS Modeler a un database Oracle mediante l'opzione Programma di caricamento esterno nella finestra di dialogo Esportazione DB: Opzioni avanzate.

Assicurarsi che il programma di utilità Oracle *sqlldr* sia installato

Lo script *oracle_loader.py* richiama il programma di utilità Oracle *sqlldr*. Si noti che *sqlldr* non viene automaticamente incluso in Oracle Client. Assicurarsi che *sqlldr* sia installata sul server su cui deve essere eseguito *oracle_loader.py*.

Specificare il SID o il nome di servizio del database

Se si stanno esportando dati in un server Oracle non locale o se il server Oracle locale ha più database, sarà necessario specificare quanto segue nel campo **Opzioni programma di caricamento esterno** nella finestra di dialogo Esportazione DB: Opzioni avanzate per trasmettere il SID o il nome di servizio:

-database <SID>

Modifica della sezione di configurazione in `oracle_loader.py`

Nei sistemi UNIX (e facoltativamente in Windows), modificare la sezione di configurazione all'inizio dello script `oracle_loader.py`. Qui è possibile specificare i valori per le variabili di ambiente `ORACLE_SID`, `NLS_LANG`, `TNS_ADMIN` e `ORACLE_HOME`, se appropriato, nonché il percorso completo del programma di utilità `sqlldr`.

Dati in formato Data, Ora e Timestamp

Nelle proprietà del flusso, si dovrebbe normalmente impostare il formato della data su **AAAA-MM-GG** e il formato dell'ora su **HH:MM:SS**.

Se si necessitano formati di data e ora diversi da quelli sopra specificati, consultare la documentazione Oracle e modificare il file di script `oracle_loader.py`.

Codifica dei dati in caratteri non ASCII

Se si stanno caricando in massa dati che non sono in formato ASCII, accertarsi che la variabile di ambiente `NLS_LANG` sia configurata correttamente nel sistema. Questa viene letta dal programma di utilità del programma di caricamento Oracle `sqlldr`. Ad esempio, il valore corretto per `NLS_LANG` per Shift-JIS in Windows è `Japanese_Japan.JA16SJIS`. Per maggiori dettagli su `NLS_LANG`, consultare la documentazione Oracle.

Stringhe vuote

Le stringhe vuote sono esportate nel database come valori NULL.

Caricamento di massa dei dati nei database SQL Server

I seguenti punti potrebbero essere d'aiuto nella configurazione del caricamento di massa da IBM SPSS Modeler a un database SQL Server mediante l'opzione Programma di caricamento esterno nella finestra di dialogo Esportazione DB: Opzioni avanzate.

Assicurarsi che il programma di utilità SQL Server `bcp.exe` sia installato

Lo script `mssql_loader.py` richiama il programma di utilità SQL Server `bcp.exe`. Assicurarsi che `bcp.exe` sia installata sul server su cui deve essere eseguito `mssql_loader.py`.

L'utilizzo degli spazi come delimitatore non funziona

Evitare di scegliere lo spazio come delimitatore nella finestra di dialogo Esportazione DB: Opzioni avanzate.

Opzione Controlla dimensione tabella consigliata

Si consiglia di abilitare l'opzione **Controlla dimensione tabella** nella finestra di dialogo Esportazione DB: Opzioni avanzate. Poiché non sempre gli errori nel processo di caricamento di massa vengono rilevati, l'abilitazione di questa opzione controllerà ulteriormente che sia stato caricato il numero corretto di righe.

Stringhe vuote

Le stringhe vuote sono esportate nel database come valori NULL.

Caricamento di massa dei dati in database Teradata

I seguenti punti potrebbero essere d'aiuto nella configurazione del caricamento di massa da IBM SPSS Modeler a un database Teradata mediante l'opzione Programma di caricamento esterno nella finestra di dialogo Esportazione DB: Opzioni avanzate.

Assicurarsi che il programma di utilità Teradata *fastload* sia installato

Lo script *teradata_loader.py* richiama il programma di utilità Teradata *fastload*. Assicurarsi che *fastload* sia installata e configurata correttamente sul server su cui deve essere eseguito *teradata_loader.py*.

I dati possono essere caricati in massa solo in tabelle vuote

È possibile utilizzare solo tabelle vuote come destinazioni per un caricamento di massa. Se una tabella di destinazione contiene dati precedenti al caricamento di massa, l'operazione non avrà successo.

Dati in formato Data, Ora e Timestamp

Nelle proprietà del flusso, impostare il formato della data su **AAAA-MM-GG** e il formato dell'ora su **HH:MM:SS**.

Stringhe vuote

Le stringhe vuote sono esportate nel database come valori NULL.

ID processo Teradata (tdpid)

Per default, *fastload* esporta i dati nel sistema Teradata con `tdpid=dbc`. Generalmente, è presente una voce nel file HOSTS che associa `dbccop1` all'indirizzo IP del server Teradata. Per utilizzare un server differente, specificare quanto riportato di seguito nel campo **Opzioni programma di caricamento esterno** nella finestra di dialogo Esportazione database - Opzioni avanzate per passare il `tdpid` di questo server:

```
-tdpid <id>
```

Spazi nei nomi di tabella e colonna

Se i nomi di tabella o di colonna contengono spazi, l'operazione di caricamento di massa non avrà successo. Se possibile, rinominare i nomi di tabella o colonna in modo da rimuovere gli spazi.

Sviluppo di programmi di caricamento di massa

Questo argomento spiega come sviluppare un programma di caricamento di massa che può essere eseguito da IBM SPSS Modeler per caricare i dati da un file di testo in un database.

Utilizzo di Python per creare programmi di caricamento di massa

Di default, IBM SPSS Modeler cerca un programma di caricamento di massa in base al tipo di database. Consultare Tabella 53 a pagina 325.

È possibile utilizzare lo script *test_loader.py* a supporto dello sviluppo di programmi di caricamento in batch. Per ulteriori informazioni, consultare l'argomento "Test dei programmi di caricamento di massa" a pagina 330.

Oggetti trasmessi al programma di caricamento di massa

IBM SPSS Modeler scrive due file che sono trasmessi al programma di caricamento di massa.

- **File dati.** Contiene i dati da caricare, in formato testo.

- **File schema.** Si tratta di un file XML che descrive i nomi e i tipi delle colonne e contiene informazioni su come è formattato il file di dati (ad esempio, quale carattere è utilizzato come delimitatore tra i campi).

Inoltre, IBM SPSS Modeler trasmette altre informazioni, quali il nome della tabella, il nome utente e la password, come argomenti quando richiama il programma di caricamento di massa.

Nota: per indicare il corretto completamento a IBM SPSS Modeler, il programma di caricamento di massa deve eliminare il file schema.

Argomenti trasmessi al programma di caricamento di massa

Gli argomenti passati al programma sono riportati nella seguente tabella.

Tabella 54. Argomenti trasmessi al programma di caricamento di massa.

Argomento	Descrizione
schemafile	Percorso del file schema.
data file	Percorso del file di dati.
servername	Nome del server DBMS; potrebbe essere vuoto.
databasename	Nome del database all'interno del server DBMS; potrebbe essere vuoto.
username	Nome utente per accedere al database.
password	Password per accedere al database.
tablename	Nome della tabella da caricare.
ownername	Nome del titolare della tabella (noto anche come nome schema).
logfile	Nome del file di registro (se vuoto, non viene creato alcun file di registro).
rowcount	Numero di righe nel dataset.

Qualsiasi opzione specificata nel campo **Opzioni programma di caricamento esterno** nella finestra di dialogo Esportazione DB: Opzioni avanzate viene passata al programma di caricamento di massa dopo questi argomenti standard.

Formato del file di dati

I dati sono scritti nel file di dati in formato testo; ogni campo viene separato con un carattere di delimitazione specificato nella finestra di dialogo Esportazione DB: Opzioni avanzate. Di seguito si riporta un esempio di come potrebbe apparire un file di dati delimitato da tabulazione.

```
48 F HIGH NORMAL 0.692623 0.055369 drugA
15 M NORMAL HIGH 0.678247 0.040851 drugY
37 M HIGH NORMAL 0.538192 0.069780 drugA
35 F HIGH HIGH 0.635680 0.068481 drugA
```

Il file è scritto nella codifica locale utilizzata da IBM SPSS Modeler Server (o da IBM SPSS Modeler se non connesso a IBM SPSS Modeler Server). Alcuni elementi della formattazione sono controllati tramite le impostazioni del flusso IBM SPSS Modeler.

Formato del file schema

Il file schema è un file XML che descrive il file di dati. Di seguito si riporta un esempio come accompagnamento al file di dati precedente.

```
<?xml version="1.0" encoding="UTF-8"?>
<DBSCHEMA version="1.0">
  <table delimiter="\t" commit_every="10000" date_format="YYYY-MM-DD" time_format="HH:MM:SS"
```

```

append_existing="false" delete_datafile="false">
  <column name="Age" encoded_name="416765" type="integer"/>
  <column name="Sex" encoded_name="536578" type="char" size="1"/>
  <column name="BP" encoded_name="4250" type="char" size="6"/>
  <column name="Cholesterol" encoded_name="43686F6C65737465726F6C" type="char" size="6"/>
  <column name="Na" encoded_name="4E61" type="real"/>
  <column name="K" encoded_name="4B" type="real"/>
  <column name="Drug" encoded_name="44727567" type="char" size="5"/>
</table>
</DBSCHEMA>

```

Le due tabelle riportate di seguito elencano gli attributi degli elementi <table> e <column> del file schema.

Tabella 55. Attributi dell'elemento <table>.

Attributo	Descrizione
delimiter	Il carattere di delimitazione del campo (il carattere TAB è rappresentato da \t).
commit_every	L'intervallo dimensione batch (come indicato nella finestra di dialogo Esportazione DB: Opzioni avanzate).
date_format	Il formato utilizzato per rappresentare le date.
time_format	Il formato utilizzato per rappresentare le ore.
append_existing	true se la tabella da caricare già contiene dati; in caso contrario, false.
delete_datafile	true se il programma di caricamento di massa deve eliminare il file dati al completamento del caricamento.

Tabella 56. Attributi dell'elemento <column>.

Attributo	Descrizione
name	Il nome della colonna.
encoded_name	Il nome della colonna convertito nella stessa codifica del file di dati e reso come una serie di numeri esadecimali a due cifre.
type	Il tipo di dati della colonna: deve essere uno tra integer, real, char, time, date e datetime.
size	Per il tipo di dati char, la larghezza massima della colonna in caratteri.

Test dei programmi di caricamento di massa

È possibile eseguire il test del caricamento di massa utilizzando un script di test *test_loader.py* incluso nella cartella *\scripts* della directory di installazione di IBM SPSS Modeler. Questa operazione è utile durante il tentativo di sviluppare, eseguire il debug o risolvere i problemi relativi agli script o ai programmi di caricamento di massa da utilizzare con IBM SPSS Modeler.

Per utilizzare lo script di test, procedere come descritto di seguito.

1. Eseguire lo script *test_loader.py* per copiare i file schema e i file di dati nei file *schema.xml* e *data.txt* e creare un file batch di Windows (*test.bat*).
2. Modificare il file *test.bat* in modo da selezionare il programma o lo script di caricamento di massa da testare.
3. Eseguire *test.bat* dalla shell comandi per testare il programma o lo script di caricamento di massa scelto.

Nota: l'esecuzione di *test.bat* non carica effettivamente i dati nel database.

Nodo di esportazione File flat

Il nodo di esportazione File flat consente di scrivere i dati in un file di testo delimitato. Questa funzione è utile per esportare i dati che possono essere letti da altri software di analisi o fogli di calcolo.

Nota: non è possibile scrivere i file nel formato della cache precedente, perché IBM SPSS Modeler non utilizza più tale formato per i file della cache. I file della cache di IBM SPSS Modeler vengono ora salvati nel formato *.sav* di IBM SPSS Statistics, che è possibile scrivere utilizzando un nodo Esporta Statistics. Per ulteriori informazioni, consultare l'argomento "Nodo Esporta Statistics" a pagina 349.

Scheda di esportazione del nodo File flat

File di esportazione. Specificare il nome del file. Digitare un nome di file o fare clic sul pulsante di selezione dei file per individuare la posizione del file.

Modalità di scrittura. Se è selezionata l'opzione **Sovrascrivi**, tutti i dati esistenti nel file specificato verranno sovrascritti. Se è selezionata l'opzione **Accoda**, l'output verrà aggiunto alla fine del file esistente, preservando tutti i dati in esso contenuti.

- **Includi nomi campi.** Se questa opzione è selezionata, i nomi dei campi verranno scritti nella prima riga del file di output. Questa opzione è disponibile solo per la modalità di scrittura **Sovrascrivi**.

Nuova riga dopo ogni record. Se questa opzione è selezionata, ogni record verrà scritto su una nuova riga nel file di output.

Separatore di campo. Specificare il carattere da inserire tra i valori di campo nel file di testo generato. Le opzioni sono **Virgola**, **Tabulatore**, **Spazio** e **Altro**. Se si seleziona **Altro**, immettere il carattere o i caratteri di delimitazione desiderati nella casella di testo.

Virgolette per i simboli. Specificare il tipo di virgolette da utilizzare per i valori dei campi simbolici. Le opzioni sono **Nessuna** (i valori non vengono racchiusi tra virgolette), **Singole (')**, **Doppie (")** e **Altro**. Se si seleziona **Altro**, immettere i caratteri desiderati nella casella di testo.

Codifica. Specifica il metodo di codifica del testo utilizzato. È possibile scegliere tra l'impostazione di default del sistema, l'impostazione di default del flusso o UTF-8.

- L'impostazione di default del sistema è specificata nel Pannello di controllo di Windows oppure, se si esegue il programma in modalità distribuita, sul computer server.
- L'impostazione di default del flusso è specificata nella finestra di dialogo Proprietà flusso.

Simbolo decimale. Specificare il tipo di separatore desiderato per i decimali nei dati.

- **Default flusso** Verrà utilizzato il separatore decimale definito dall'impostazione di default del flusso corrente. In genere si tratterà del separatore decimale definito dalle impostazioni internazionali del computer.
- **Punto (.)**. Come separatore di decimali verrà utilizzato il punto.
- **Virgola (,)**. Come separatore di decimali verrà utilizzata la virgola.

Genera un nodo di importazione per questi dati. Selezionare questa opzione per generare automaticamente un nodo origine Testo variabile che leggerà il file di dati esportato. Per ulteriori informazioni, consultare l'argomento "Nodo Testo variabile" a pagina 20.

Nodo Esporta IBM SPSS Data Collection

Il nodo Esportazione di IBM SPSS Data Collection salva i dati nel formato utilizzato dai prodotti per ricerche di mercato IBM SPSS Data Collection, in base al Modello dati di IBM SPSS Data Collection. Questo formato distingue i dati del caso — le risposte effettive alle domande raccolte durante un'indagine — dai metadati che descrivono il modo in cui i dati del caso vengono raccolti ed organizzati.

I metadati sono costituiti da informazioni quali testi di domande, nomi di variabili e descrizioni, insiemi di risposte multiple, traduzioni dei vari testi, nonché dalla definizione della struttura dei dati del caso. Per ulteriori informazioni, consultare l'argomento "Nodo Data Collection" a pagina 25.

Nota: questo nodo richiede IBM SPSS Data Collection Data Model versione 4.0 o versioni successive, distribuito con il software IBM SPSS Data Collection. Per ulteriori informazioni, vedere la pagina Web di IBM SPSS Data Collection all'indirizzo <http://www.ibm.com/software/analytics/spss/products/data-collection/>. Fatta eccezione per l'installazione di Data Model, non è necessario eseguire altre operazioni di configurazione.

File di metadati. Specifica il nome del file di definizione del questionario (*.mdd*) in cui verranno salvati i metadati esportati. Viene creato un questionario di default in base alle informazioni sui tipi di campi. Per esempio, un campo nominale (insieme) potrebbe essere rappresentato come una domanda singola con la descrizione del campo utilizzata come testo della domanda e una casella di controllo separata per ogni valore definito.

Unisci metadati. Specifica se i metadati sovrascriveranno le versioni esistenti o saranno uniti ai metadati esistenti. Se si seleziona l'opzione di unione, ogni volta che viene eseguito il flusso viene creata una nuova versione. In questo modo è possibile tenere traccia delle versioni di un questionario via via che questo viene modificato. Ogni versione può essere considerata come una snapshot dei metadati utilizzati per raccogliere un particolare insieme di dati del caso.

Attiva variabili di sistema. Specifica se le variabili di sistema vengono incluse nel file *.mdd* esportato. Tra queste sono comprese variabili come *Respondent.Serial*, *Respondent.Origin* e *DataCollection.StartTime*.

Impostazioni dei dati del caso. Specifica il file di dati IBM SPSS Statistics (*.sav*) in cui vengono esportati i dati del caso. Si noti che in questo caso sono valide tutte le restrizioni sui nomi delle variabili e dei valori; quindi, per esempio, può essere necessario passare alla scheda Filtro e utilizzare l'opzione "Rinomina per IBM SPSS Statistics" del menu Filtro per correggere i caratteri non validi nei nomi dei campi.

Genera un nodo di importazione per questi dati. Selezionare questa opzione per generare automaticamente un nodo di origine IBM SPSS Data Collection che leggerà il file di dati esportato.

Insiemi di risposte multiple. Tutti gli insiemi di risposte multiple definiti nel flusso saranno automaticamente conservati quando il file viene esportato. Gli insiemi di risposte multiple si possono visualizzare e modificare da qualsiasi nodo mediante la scheda Filtro. Per ulteriori informazioni, consultare l'argomento "Modifica di insiemi di risposte multiple" a pagina 124.

Esportazione di Analytic Server

L'esportazione di Analytic Server consente di scrivere i dati dall'analisi ad un'origine dati Analytic Server esistente. Può trattarsi, ad esempio, di file di testo su HDFS (Hadoop Distributed File System) o di un database.

In genere, anche un flusso con nodo di esportazione Analytic Server inizia con i nodi Origine Analytic Server e viene inoltrato a Analytic Server ed eseguito su HDFS. In alternativa, un flusso con origini dati "locali" può terminare con un nodo di esportazione Analytic Server per poter caricare dataset relativamente piccoli (non più di 100.000 record) da utilizzare con Analytic Server.

Origine dati. Selezionare un'origine dati che contiene i dati che si desidera utilizzare. Un'origine dati contiene i file e i metadati associati a quella origine. Fare clic su **Seleziona** per visualizzare un elenco delle origini dati disponibili. Per ulteriori informazioni, consultare l'argomento "Seleziona origine dati di Analytic Server" a pagina 30.

Se è necessario creare una nuova origine dati o modificarne una esistente, fare clic su **Avvia editor origine dati...**

Modalità. Selezionare **Accoda** per l'aggiunta all'origine dati esistente o **Sovrascrivi** per sostituire il contenuto dell'origine dati.

Genera un nodo di importazione per questi dati. Selezionare questa opzione per generare un nodo origine per i dati esportati nell'origine dati specificata. Questo nodo viene aggiunto all'area di disegno del flusso.

Nodo di esportazione IBM Cognos BI

Il nodo di esportazione IBM Cognos BI consente di esportare dati da un flusso di IBM SPSS Modeler a Cognos BI, in formato UTF-8. In questo modo, Cognos BI può usufruire dei dati trasformati o dei punteggi dei dati di IBM SPSS Modeler. Per esempio, si può utilizzare Cognos BI Report Studio per creare un report basato sui dati esportati, comprese le previsioni e i valori di confidenza. Il report si può in seguito salvare sul server Cognos BI ed essere distribuito agli utenti di Cognos BI.

Nota: è possibile esportare solo dati relazionali, non dati OLAP.

Per esportare i dati in Cognos BI è necessario specificare quanto segue:

- Connessione Cognos - la connessione al server Cognos BI
- Connessione ODBC - la connessione al server di dati Cognos utilizzato dal server Cognos BI

All'interno della connessione Cognos si specifica una sorgente dati Cognos da utilizzare. Questa sorgente dati deve utilizzare lo stesso login della sorgente dati ODBC.

I dati del flusso vengono esportati sul server di dati, mentre i metadati del pacchetto vengono esportati sul server Cognos BI.

Come con qualsiasi altro nodo di esportazione, è possibile anche utilizzare la scheda Pubblica della finestra di dialogo del nodo per pubblicare il flusso per la distribuzione mediante IBM SPSS Modeler Solution Publisher.

Connessione Cognos

Qui è possibile specificare la connessione al server Cognos BI da utilizzare per l'esportazione. La procedura prevede l'esportazione dei metadati in un nuovo pacchetto sul server Cognos BI, mentre i dati del flusso vengono esportati sul server di dati Cognos.

Connessione. Fare clic sul pulsante **Modifica** per visualizzare una finestra di dialogo in cui è possibile definire l'URL ed altri dettagli del server Cognos BI in cui si desidera esportare i dati. Se si è già connessi a un server Cognos BI mediante IBM SPSS Modeler è possibile anche modificare i dettagli della connessione corrente. Per ulteriori informazioni, consultare l'argomento "Connessioni Cognos" a pagina 37.

Origine dati. Il nome della sorgente dati Cognos (in genere un database) in cui saranno esportati i dati. L'elenco a discesa mostra tutte le sorgenti dati Cognos a cui è possibile accedere dalla connessione corrente. Fare clic sul pulsante **Aggiorna** per aggiornare l'elenco.

Cartella. Il percorso e il nome della cartella del server Cognos BI dove deve essere creato il pacchetto di esportazione.

Nome pacchetto. Il nome del pacchetto che dovrà contenere i metadati esportati all'interno della cartella specificata. Deve essere un nuovo pacchetto con un solo oggetto della query; non è possibile esportare i metadati in un pacchetto esistente.

Modalità. Specifica le modalità di esecuzione dell'esportazione:

- **Pubblica subito il pacchetto.** (impostazione di default) Esegue l'operazione di esportazione non appena si fa clic su **Esegui**.
- **Script di esportazione.** Crea uno script XML di esportazione eseguibile in un secondo momento (per esempio tramite Framework Manager). Digitare il percorso e il nome del file di script nel campo **File**, oppure specificare il nome e la posizione del file di script con il pulsante **Modifica**.

Genera un nodo di importazione per questi dati. Selezionare questa opzione per generare un nodo origine per i dati esportati nella sorgente dati e nella tabella specificate. Quando si fa clic su **Esegui**, questo nodo viene aggiunto all'area del flusso.

Connessione ODBC

Qui viene specificata la connessione al server di dati Cognos (cioè al database) in cui saranno esportati i dati del flusso.

Nota: è necessario verificare che l'origine dati specificata qui punti alla stessa origine dati specificata nel pannello **Connessioni Cognos**. È inoltre necessario assicurarsi che la sorgente dati della connessione Cognos utilizzi lo stesso login della sorgente dati ODBC.

Origine dati. Mostra la sorgente dati selezionata. Immettere il nome o selezionarlo dall'elenco a discesa. Se nell'elenco non compare il database desiderato, selezionare **Aggiungi nuova connessione database** e individuare il database dalla finestra di dialogo Connessioni database. Per ulteriori informazioni, consultare l'argomento "Aggiunta di una connessione a un database" a pagina 15.

Nome tabella. Immettere il nome della tabella alla quale si desidera inviare i dati. Se si seleziona l'opzione **Inserisci in tabella**, è possibile selezionare una tabella esistente del database facendo clic sul pulsante **Seleziona**.

Crea tabella. Selezionare questa opzione per creare una nuova tabella di database o per sovrascriverne una esistente.

Inserisci in tabella. Selezionare questa opzione per inserire i dati come nuove righe in una tabella di database esistente.

Unisci tabella. (Dove disponibile) Selezionare questa opzione per aggiornare le colonne del database selezionate con i valori dei campi dati di origine corrispondenti. La selezione di questa opzione abilita il pulsante **Unione**, che visualizza una finestra di dialogo da cui è possibile mappare i campi dati di origine alle colonne del database.

Elimina tabella esistente. Selezionare questa opzione per eliminare qualsiasi tabella esistente con lo stesso nome della nuova tabella che si desidera creare.

Elimina righe esistenti. Selezionare questa opzione per eliminare dalla tabella le righe esistenti prima dell'esportazione quando si inseriscono dati in una tabella.

Nota: se è selezionata una delle due opzioni sopra riportate, durante l'esecuzione del nodo verrà visualizzato un messaggio **Avviso di sovrascrittura**. Per eliminare gli avvisi, deselezionare **Avvisa quando un nodo sovrascrive una tabella di database** nella scheda Notifiche della finestra di dialogo Opzioni utente.

Dimensioni di default stringhe. I campi contrassegnati come senza tipo in un nodo Tipo a monte vengono scritti nel database come campi stringa. Specificare la dimensione delle stringhe da utilizzare per i campi senza tipo.

Fare clic su **Schema** per visualizzare una finestra di dialogo in cui è possibile configurare diverse opzioni di esportazione (per i database che supportano tale funzione), impostare i tipi di dati SQL per i campi e specificare la chiave primaria ai fini dell'indicizzazione del database. Per ulteriori informazioni, consultare l'argomento "Esportazione Database - Opzioni Schema" a pagina 318.

Fare clic su **Indici** per impostare le opzioni di indicizzazione della tabella esportata al fine di migliorare le prestazioni del database. Per ulteriori informazioni, consultare l'argomento "Esportazione database - Opzioni Indice" a pagina 321.

Fare clic su **Opzioni avanzate** per specificare le opzioni di caricamento di massa e di commit nel database. Per ulteriori informazioni, consultare l'argomento "Esportazione database - Opzioni avanzate" a pagina 323.

Nomi di tabelle e colonne fra virgolette. Selezionare le opzioni da utilizzare al momento dell'invio di un'istruzione CREATE TABLE al database. Le tabelle e le colonne con spazi o caratteri non standard devono essere racchiuse tra virgolette.

- **Se necessario.** Selezionare questa opzione per consentire a IBM SPSS Modeler di determinare automaticamente l'eventuale esigenza di virgolette per ogni singolo caso.
- **Sempre.** Selezionare questa opzione per racchiudere sempre i nomi di tabella e di colonna tra virgolette.
- **Mai.** Selezionare questa opzione per disattivare l'utilizzo delle virgolette.

Genera un nodo di importazione per questi dati. Selezionare questa opzione per generare un nodo origine per i dati esportati nella sorgente dati e nella tabella specificate. Quando si fa clic su **Esegui**, questo nodo viene aggiunto all'area del flusso.

Nodo di esportazione IBM Cognos TM1

Il nodo di esportazione IBM Cognos BI consente di esportare i dati da un flusso IBM SPSS Modeler in Cognos TM1. In questo modo, Cognos BI può usufruire dei dati trasformati o dei punteggi dei dati di IBM SPSS Modeler.

Nota: è possibile esportare solo misure non dati di dimensione contesto; in alternativa, è possibile aggiungere nuovi elementi al cubo.

Per esportare i dati in Cognos BI è necessario specificare quanto segue:

- La connessione al server Cognos TM1
- Il cubo in cui verranno esportati i dati
- Il mapping dei nomi di dati SPSS alle dimensioni e misure TM1 equivalenti

Come con qualsiasi altro nodo di esportazione, è possibile anche utilizzare la scheda Pubblica della finestra di dialogo del nodo per pubblicare il flusso per la distribuzione mediante IBM SPSS Modeler Solution Publisher.

Connessione a un cubo Cognos TM1 per esportare i dati

La prima fase nell'esportazione dei dati in un database IBM Cognos TM1 è selezionare l'hub PM TM1 relativo e server e cubo associati, nella scheda **Connessione** della finestra di dialogo di IBM Cognos TM1.

Nota: È necessario assicurarsi che la struttura dati della vista su cui si sta effettuando l'esportazione corrisponda alla struttura dati della vista da cui sono stati importati i dati. La vista utilizzata per l'esportazione può essere su un cubo diverso rispetto a quello utilizzato per l'importazione purché entrambe le viste condividano la stessa struttura.

Sistema PM. Immettere l'URL dell'hub che contiene il server TM1 a cui ci si desidera **Connettere**.

Server TM1. Dopo aver stabilito la connessione all'hub Cognos, selezionare il server che contiene i dati da importare e fare clic su **Login**. Se non ci si è mai connessi in precedenza a questo server viene richiesta l'immissione del **Nome utente** e della **Password**; in alternativa, è possibile selezionare un altro server.

Selezionare un cubo TM1 da esportare. Visualizza il nome dei cubi all'interno del server TM1 in cui è possibile esportare i dati. Fare doppio clic su un cubo per visualizzare le viste contenute in quel cubo.

Per scegliere i dati da esportare, selezionare il cubo e fare clic sulla freccia destra per spostare il cubo nel campo **Esporta nel cubo**. Dopo aver selezionato il cubo, utilizzare la scheda **Mapping** per associare i campi TM1 ai campi SPSS relativi.

Mapping dei dati Cognos TM1 per l'esportazione

Dopo aver selezionato l'hub PM del TM1 e il server e il cubo associati, nella scheda **Connessione** della finestra di dialogo di IBM Cognos TM, si identifica quali campi SPSS mappano i campi TM1 associati.

Dove possibile i campi TM1 vengono associati automaticamente ai campi SPSS pertinenti.

Nota: È necessario assicurarsi che la struttura dati della vista su cui si sta effettuando l'esportazione corrisponda alla struttura dati della vista da cui sono stati importati i dati. La vista utilizzata per l'esportazione può essere su un cubo diverso rispetto a quello utilizzato per l'importazione purché entrambe le viste condividano la stessa struttura.

Campi. Elenca il nome del campo dei dati, fornito dal file di dati SPSS, di dati disponibili per l'esportazione. Se richiesto, utilizzare i pulsanti in fondo all'elenco per modificare i campi mostrati; ad esempio, è possibile mostrare tutti i campi, solo quelli di un tipo continuo o quelli di tipo categorico.

Dimensioni TM1. Mostra il cubo selezionato nella scheda **Connessione**, insieme alle sue dimensioni regolari e di misura. Se non viene completato automaticamente, selezionare il nome dell'elemento TM1 da associare al campo dati SPSS.

Nodo Esporta SAS

Nota: questa funzione è disponibile in SPSS Modeler Professional e SPSS Modeler Premium.

Il nodo **Esporta SAS** consente di scrivere i dati in formato SAS, in modo da essere letti in SAS o un package di software compatibile con SAS. È possibile eseguire l'esportazione in tre formati di file SAS: SAS per Windows/OS2, SAS per UNIX o SAS Versione 7/8.

Scheda Esporta del nodo Esporta SAS

File di esportazione. Specificare il nome del file. Digitare un nome di file o fare clic sul pulsante di selezione dei file per individuare la posizione del file.

Esporta. Specificare il formato del file di esportazione. Le opzioni sono **SAS per Windows/OS2**, **SAS per UNIX** o **SAS Versione 7/8**.

Esporta nomi di campo come variabile. Selezionare le opzioni per esportare i nomi e le etichette di campo da IBM SPSS Modeler per utilizzarli in SAS.

- **Nomi ed etichette di variabile.** Selezionare questa opzione per esportare sia i nomi che le etichette dei campi di IBM SPSS Modeler. I nomi vengono esportati come nomi di variabili SAS, mentre le etichette vengono esportate come etichette di variabile SAS.

- **Nomi come etichette di variabile.** Selezionare questa opzione per utilizzare i nomi di campi IBM SPSS Modeler come etichette di variabile SAS. IBM SPSS Modeler consente di includere nei nomi dei campi caratteri non validi nei nomi di variabili SAS. Per evitare di creare nomi SAS non validi, selezionare **Nomi ed etichette di variabile**.

Genera un nodo di importazione per questi dati. Selezionare questa opzione per generare automaticamente un nodo origine SAS che leggerà il file di dati esportato. Per ulteriori informazioni, consultare l'argomento "Nodo origine SAS" a pagina 39.

Nodo Esportazione da Excel

Il nodo Esportazione da Excel esegue l'output di dati in formato Microsoft Excel (.xls). Se lo si desidera, è possibile scegliere di avviare Excel automaticamente e aprire il file esportato quando si esegue il nodo.

Scheda di esportazione del nodo Excel

Nome file. Digitare un nome di file o fare clic sul pulsante di selezione dei file per individuare la posizione del file. Il nome file di default è *excelexp.xls*.

Tipo di file. Selezionare il tipo di file Excel che si sta esportando.

Crea nuovo file. Crea un nuovo file di Excel.

Inserisci nel file esistente. Il contenuto viene sostituito a partire dalla cella specificata nel campo **Avvia nella cella**. Le altre celle del foglio di calcolo mantengono il contenuto originale.

Includi nomi campi. Specifica se i nomi dei campi devono essere inclusi nella prima riga del foglio di lavoro.

Avvia nella cella. La posizione della cella utilizzata per il primo record di esportazione (o il nome del primo campo se è selezionato **Includi nomi campi**). I dati vengono compilati a destra e verso il basso rispetto a questa cella iniziale.

Scegli foglio di lavoro. Specifica il foglio di lavoro nel quale si desidera esportare i dati. È possibile individuare il foglio di lavoro in base all'indice o al nome:

- **Tramite indice.** In caso di creazione di un nuovo file, specificare un valore tra 0 e 9 per identificare il foglio di lavoro nel quale si desidera effettuare l'esportazione, iniziando con 0 per il primo foglio di lavoro, 1 per il secondo e così via. È possibile utilizzare valori maggiori o uguali a 10 solo se in questa posizione esiste già un foglio di lavoro.
- **Per nome.** In caso di creazione di un nuovo file, specificare il nome utilizzato per il foglio di lavoro. In caso di inserimento all'interno di un file esistente, se il foglio di lavoro esiste i dati vengono inseriti al suo interno, altrimenti viene creato un nuovo foglio di lavoro con questo nome.

Avvia Excel. Indica se Excel deve essere avviato automaticamente con il file esportato quando viene eseguito il nodo. Si noti che, durante l'esecuzione in modalità distribuita con IBM SPSS Modeler Server, l'output viene salvato nel file system del server ed Excel viene avviato nell'installazione Client con una copia del file esportato.

Genera un nodo di importazione per questi dati. Selezionare questa opzione per generare automaticamente un nodo origine Excel che leggerà il file di dati esportato. Per ulteriori informazioni, consultare l'argomento "Nodo origine Excel" a pagina 40.

Nodo Esporta XML

Il nodo Esporta XML consente di esportare i dati in formato XML, utilizzando la codifica UTF-8. Se lo si desidera, è possibile creare un nodo origine XML per leggere nuovamente i dati esportati nel flusso.

File di esportazione XML. Il percorso e il nome file completi del file XML nel quale si desidera esportare i dati.

Utilizza schema XML. Selezionare questa casella di controllo per utilizzare uno schema o DTD per controllare la struttura dei dati esportati. In questo modo viene attivato il pulsante **Mappa** descritto di seguito.

Se non si utilizza uno schema o un DTD, per i dati esportati viene utilizzata la seguente struttura di default:

```
<records>
  <record>
    <fieldname1>value</fieldname1>
    <fieldname2>value</fieldname2>
    :
    <fieldnameN>value</fieldnameN>
  </record>
  <record>
    :
    :
  </record>
  :
  :
</records>
```

Gli spazi in un nome campo vengono sostituiti dai caratteri di sottolineatura, per esempio "My Field" diventa <My_Field>.

Mappa. Se è stata selezionata l'opzione di utilizzo di uno schema XML, il pulsante visualizza una finestra di dialogo nella quale è possibile specificare la parte della struttura XML da utilizzare per iniziare ogni nuovo record. Per ulteriori informazioni, consultare l'argomento "Opzioni dei record di mappatura XML" a pagina 339.

Campi mappati. Indica il numero di campi che sono stati mappati.

Genera un nodo di importazione per questi dati. Selezionare questa opzione per generare automaticamente un nodo origine XML che rileggerà il file di dati esportato nel flusso. Per ulteriori informazioni, consultare l'argomento "Nodo origine XML" a pagina 41.

Scrittura dei dati XML

Quando viene specificato un elemento XML, il valore del campo viene posizionato all'interno del tag dell'elemento:

```
<element>value</element>
```

Quando viene mappato un attributo, il valore del campo viene posizionato come valore dell'attributo:

```
<element attribute="value">
```

Se un campo viene mappato a un elemento sopra all'elemento <records>, il campo viene scritto solo una volta e sarà una costante per tutti i record. Il valore di questo elemento proviene dal primo record.

Se è necessario scrivere un valore null, specificare un contenuto vuoto. Per gli elementi, si tratta di:

```
<element></element>
```

Per gli attributi, si tratta di:

```
<element attribute="">
```

Opzioni dei record di mappatura XML

La scheda Record consente di specificare quale parte della struttura XML debba essere utilizzata per iniziare ogni nuovo record. Per eseguire correttamente la mappatura a uno schema, è necessario specificare il delimitatore dei record.

Struttura XML. Una struttura ad albero gerarchica che mostra la struttura dello schema XML specificato nella schermata precedente.

Record (espressione XPath). Per impostare il delimitatore dei record, selezionare un elemento nella struttura XML e fare clic sul pulsante con la freccia destra. Ogni volta che si incontra questo elemento nei dati di input, nel file di output viene creato un nuovo record.

Nota: se nella struttura XML viene selezionato l'elemento root, è possibile scrivere solo un record singolo e tutti gli altri record vengono ignorati.

Opzioni dei campi di mappatura XML

La scheda Campi viene utilizzata per mappare i campi nell'insieme di dati a elementi o attributi nella struttura XML quando viene utilizzato un file schema.

I nomi dei campi che corrispondono al nome di un elemento o di un attributo vengono mappati automaticamente a condizione che il nome dell'elemento o dell'attributo sia univoco. Pertanto, se sono presenti sia un elemento che un attributo con il nome `field1`, non viene eseguita alcuna mappatura automatica. Se nella struttura è presente un solo elemento di nome `field1`, viene mappato automaticamente un campo con questo nome nel flusso.

Campi. L'elenco dei campi nel modello. Selezionare uno o più campi come parte di input della mappatura. È possibile utilizzare i pulsanti in fondo all'elenco per selezionare tutti i campi, oppure tutti i campi con un particolare livello di misurazione.

Struttura XML. Selezionare un elemento nella struttura XML come obiettivo della mappatura. Per generare la mappatura, fare clic su Mappa. La mappatura viene quindi visualizzata. Il numero di campi mappati in questo modo viene visualizzato sotto all'elenco.

Per eliminare una mappatura, selezionare l'elemento nell'elenco della struttura XML e fare clic su **Annulla mappatura**.

Visualizza attributi. Visualizza o nasconde gli attributi, se presenti, degli elementi XML nella struttura XML.

Anteprima mappatura XML

Nella scheda Anteprima, fare clic su **Aggiorna** per visualizzare un'anteprima dell'XML che verrà scritto.

Se la mappatura non è corretta, ritornare alla scheda Record o Campi per correggere gli errori e fare nuovamente clic su **Aggiorna** per visualizzare il risultato.

Capitolo 8. Nodi IBM SPSS Statistics

Nodi IBM SPSS Statistics - Panoramica

Come complemento a IBM SPSS Modeler e alle relative funzionalità di data mining, IBM SPSS Statistics consente di eseguire analisi statistiche più approfondite e di gestire i dati.

Se si dispone di una copia compatibile con licenza di IBM SPSS Statistics installata, è possibile connettersi a questo programma da IBM SPSS Modeler ed eseguire analisi e manipolazioni di dati complessi a più fasi che non sarebbero altrimenti supportate da IBM SPSS Modeler. L'utente più esperto ha inoltre la possibilità di modificare ulteriormente le analisi tramite una sintassi di comandi. Vedere le Note di versione per informazioni sulla compatibilità delle versioni.

Se disponibili, i nodi IBM SPSS Statistics vengono mostrati nella parte dedicata della palette dei nodi.

Nota: si consiglia di istanziare i dati in un nodo Tipo prima di utilizzare i nodi Trasformazioni, Modello o Output di IBM SPSS Statistics. Questo è necessario anche quando si utilizza il comando di sintassi AUTORECODE.

La tavolozza IBM SPSS Statistics contiene i seguenti nodi:



Il nodo File Statistics legge i dati dal formato file *.sav* o *.zsav* utilizzato da IBM SPSS Statistics, oltre ai file di cache salvati in IBM SPSS Modeler, che utilizzano lo stesso formato.



Il nodo Trasformazioni Statistics esegue una selezione di comandi di sintassi IBM SPSS Statistics rispetto alle sorgenti dati in IBM SPSS Modeler. Questo nodo richiede una copia di IBM SPSS Statistics con regolare licenza.



Il nodo Modello Statistics consente di analizzare e operare con i dati eseguendo le procedure IBM SPSS Statistics che generano PMML. Questo nodo richiede una copia di IBM SPSS Statistics con regolare licenza.



Il nodo Output Statistics consente di chiamare una procedura IBM SPSS Statistics per analizzare i dati di IBM SPSS Modeler. È disponibile una vasta gamma di procedure analitiche di IBM SPSS Statistics. Questo nodo richiede una copia di IBM SPSS Statistics con regolare licenza.



Il nodo Esporta Statistics restituisce i dati in formato IBM SPSS Statistics *.sav* o *.zsav*. I file *.sav* o *.zsav* possono essere letti da IBM SPSS Statistics Base e da altri prodotti. Questo formato viene inoltre utilizzato per i file cache di IBM SPSS Modeler.

Nota: se la propria copia di SPSS Statistics dispone di licenza solo per un utente singolo e viene eseguito un flusso con due o più rami, ciascuno dei quali contiene un nodo SPSS Statistics, è possibile che venga visualizzato un errore relativo alla licenza. Ciò si verifica quando la sessione di SPSS Statistics di un ramo

non è stata completata prima che venga avviata la sessione di un altro ramo. Se possibile, riprogettare il flusso in modo che non vengano eseguiti in parallelo più rami con nodi SPSS Statistics.

Nodo File Statistics

È possibile utilizzare il nodo File Statistics per leggere i dati direttamente da un file IBM SPSS Statistics salvato (.sav o .zsav). Questo formato viene attualmente utilizzato per sostituire il file cache di versioni precedenti di IBM SPSS Modeler. Se si desidera importare un file cache salvato, è necessario utilizzare il nodo File IBM SPSS Statistics.

File di importazione. Specificare il nome del file. È possibile immettere un nome di file o fare clic sul pulsante con i puntini di sospensione (...) per selezionare un file. Il percorso di un file viene visualizzato dopo la selezione del file.

Il file è crittografato da una password. Selezionare questa casella se si è al corrente che il file è protetto da password; verrà poi richiesto di immettere la **Password**. Se il file è protetto da password e non si immette la password, viene visualizzato un messaggio di avviso se si tenta di passare ad un'altra scheda, di aggiornare i dati, di visualizzare in anteprima il contenuto del nodo o di eseguire un flusso che contiene il nodo.

Nota: I file protetti da password possono essere aperti solo da IBM SPSS Modeler versione 16 o superiore.

Nomi delle variabili. Selezionare un metodo per la gestione dei nomi e delle etichette di variabile durante l'importazione da un file IBM SPSS Statistics .sav o .zsav. I metadati che si sceglie di includere in questo punto vengono mantenuti per tutta la sessione di lavoro in IBM SPSS Modeler e possono essere nuovamente esportati per l'utilizzo in IBM SPSS Statistics.

- **Leggi nomi ed etichette.** Selezionare questa opzione per leggere sia i nomi che le etichette di variabile in IBM SPSS Modeler. Questa opzione è selezionata per default e i nomi di variabile vengono visualizzati nel nodo Tipo. È possibile visualizzare le etichette nei grafici, nei browser dei modelli e in altri tipi di output, in base alle opzioni specificate nella finestra di dialogo delle proprietà del flusso. La visualizzazione delle etichette nell'output è disattivata per default.
- **Leggi etichette come nomi.** Selezionare questa opzione per leggere le etichette di variabile descrittive dal file IBM SPSS Statistics .sav o .zsav invece dei nomi di campo brevi ed utilizzare tali etichette come nomi di variabili in IBM SPSS Modeler.

Valori. Selezionare un metodo per la gestione dei valori e delle etichette durante l'importazione da un file IBM SPSS Statistics .sav o .zsav. I metadati che si sceglie di includere in questo punto vengono mantenuti per tutta la sessione di lavoro in IBM SPSS Modeler e possono essere nuovamente esportati per l'utilizzo in IBM SPSS Statistics.

- **Leggi dati ed etichette.** Selezionare questa opzione per leggere sia i valori effettivi che le etichette dei valori in IBM SPSS Modeler. Questa opzione è selezionata per default e i valori vengono visualizzati nel nodo Tipo. Le etichette dei valori possono essere visualizzate nel generatore di espressioni, nei grafici, nei browser dei modelli e in altri tipi di output, in base alle opzioni specificate nella finestra di dialogo delle proprietà del flusso.
- **Leggi etichette come dati.** Selezionare questa opzione se si desidera utilizzare le etichette dei valori del file .sav o .zsav anziché i codici numerici o simbolici utilizzati per rappresentare i valori. Per esempio, se si seleziona questa opzione per dati con un campo relativo al genere i cui valori 1 e 2 rappresentano rispettivamente *maschio* e *femmina*, il campo verrà convertito in una stringa e come valori effettivi verranno importati i valori *maschio* e *femmina*.

È importante considerare i valori mancanti nei dati IBM SPSS Statistics prima di selezionare questa opzione. Ad esempio, se un campo numerico utilizza le etichette solo per i valori mancanti (0 = *Nessuna risposta*, -99 = *Sconosciuto*), selezionando l'opzione precedente verranno importate solo le etichette di valori *Nessuna risposta* e *Sconosciuto* ed il campo verrà convertito in una stringa. In questi casi, è consigliabile importare i valori stessi e impostare i valori mancanti in un nodo Tipo.

Utilizza informazioni sul formato di campo per determinare l'archiviazione. Se questa casella è deselezionata, i valori dei campi formattati nel file *.sav* come numeri interi (ad esempio i campi specificati come *Fn.0* nella vista delle variabili in IBM SPSS Statistics) vengono importati utilizzando l'archiviazione numero intero. Tutti i valori degli altri campi tranne le stringhe vengono importati come numeri reali.

Se questa casella è selezionata (default), tutti i valori dei campi tranne le stringhe vengono importati come numeri reali, che siano formattati o meno come numeri interi nel file *.sav*.

Insiemi di risposte multiple. Tutti gli insiemi di risposte multiple definiti nel file IBM SPSS Statistics vengono mantenuti automaticamente quando il file viene importato. Gli insiemi di risposte multiple si possono visualizzare e modificare da qualsiasi nodo mediante la scheda Filtro. Per ulteriori informazioni, consultare l'argomento "Modifica di insiemi di risposte multiple" a pagina 124.

Nodo Trasformazioni Statistics

Il nodo Trasformazioni Statistics consente di completare le trasformazioni dati utilizzando la sintassi di comando IBM SPSS Statistics. Questa funzione consente di completare una serie di trasformazioni non supportate da IBM SPSS Modeler e di automatizzare trasformazioni complesse e articolate in più fasi, compresa la creazione di alcuni campi da un unico nodo. È simile al nodo Output Statistics, con l'eccezione che i dati vengono restituiti a IBM SPSS Modeler per ulteriori analisi, mentre nel nodo Output i dati vengono restituiti come oggetti di output richiesti, come grafici o tabelle.

Per utilizzare questo nodo è necessario avere installata sul computer una versione compatibile di IBM SPSS Statistics con regolare licenza. Per ulteriori informazioni, consultare l'argomento "IBM SPSS Statistics Applicazioni di supporto" a pagina 312. Per informazioni relative alla compatibilità, consultare le note sul rilascio.

Se necessario, è possibile visualizzare la scheda Filtro per filtrare o rinominare i campi in modo che risultino conformi agli standard di denominazione IBM SPSS Statistics. Per ulteriori informazioni, consultare l'argomento "Ridenominazione o filtro dei campi per IBM SPSS Statistics" a pagina 350.

Syntax Reference. Per informazioni dettagliate su specifiche procedure IBM SPSS Statistics, consultare la guida *IBM SPSS Statistics Command Syntax Reference*, inclusa nella copia del software di IBM SPSS Statistics. Per visualizzare la guida dalla scheda Sintassi, selezionare l'opzione **Editor di sintassi** e fare clic sul pulsante Avvia Guida sintassi IBM SPSS Statistics.

Nota: non tutta la sintassi IBM SPSS Statistics è supportata da questo nodo. Per ulteriori informazioni, consultare l'argomento "Sintassi consentita" a pagina 344.

Scheda Sintassi del nodo Trasformazioni Statistics

Opzione finestra di dialogo IBM SPSS Statistics

Se non si conosce la sintassi di IBM SPSS Statistics per una procedura, il modo più semplice per creare una sintassi in IBM SPSS Modeler consiste nello scegliere l'opzione **Finestra di dialogo IBM SPSS Statistics**, selezionare la finestra di dialogo per la procedura, compilare la finestra e fare clic su OK. La sintassi verrà inserita nella scheda Sintassi del nodo IBM SPSS Statistics che si sta utilizzando in IBM SPSS Modeler. Sarà quindi possibile eseguire il flusso per ottenere l'output dalla procedura.

Opzione Editor di sintassi IBM SPSS Statistics

Controllo. Dopo che sono stati immessi i comandi di sintassi nella parte superiore della finestra di dialogo, utilizzare questo pulsante per convalidare i dati immessi. Gli eventuali errori di sintassi vengono evidenziati nella parte inferiore della finestra.

Per evitare il prolungarsi del processo di verifica, quando viene convalidata la sintassi viene eseguita la verifica in base a un campione rappresentativo dei dati e non in base all'insieme di dati completo.

Sintassi consentita

Se si dispone di una grande quantità di sintassi legacy da IBM SPSS Statistics oppure si ha già una certa dimestichezza con le funzioni di preparazione dei dati di IBM SPSS Statistics, è possibile utilizzare il nodo Trasformazioni Statistics per eseguire molte delle trasformazioni esistenti. Come linea guida, il nodo consente di trasformare i dati in modi prevedibili, per esempio, mediante l'esecuzione di comandi in ciclo oppure mediante la modifica, l'aggiunta, l'ordinamento, il filtraggio o la selezione dei dati.

Segue un elenco di alcuni dei comandi che possono essere eseguiti:

- Calcolo di numeri random in base a una distribuzione binomiale:
`COMPUTE newvar = RV.BINOM(10000,0.1)`
- Ricodifica di una variabile in una nuova variabile:
`RECODE Age (Lowest thru 30=1) (30 thru 50=2) (50 thru Highest=3) INTO AgeRecoded`
- Sostituzione dei valori mancanti:
`RMV Age_1=SMEAN(Age)`

La sintassi IBM SPSS Statistics supportata dal nodo Trasformazioni Statistics è riportata di seguito.

Nome comando

ADD VALUE LABELS
APPLY DICTIONARY
AUTORECODE
BREAK
CD
CLEAR MODEL PROGRAMS
CLEAR TIME PROGRAM
CLEAR TRANSFORMATIONS
COMPUTE
COUNT
CREATE
DATE
DEFINE-!ENDDFINE
DELETE VARIABLES
DO IF
DO REPEAT
ELSE
ELSE IF
END CASE
END FILE
END IF
END INPUT PROGRAM
END LOOP
END REPEAT
EXECUTE

Nome comando

FILE HANDLE
FILE LABEL
FILE TYPE-END FILE TYPE
FILTER
FORMATS
IF
INCLUDE
INPUT PROGRAM-END INPUT PROGRAM
INSERT
LEAVE
LOOP-END LOOP
MATRIX-END MATRIX
MISSING VALUES
N OF CASES
NUMERIC
PERMISSIONS
PRESERVE
RANK
RECODE
RENAME VARIABLES
RESTORE
RMV
SAMPLE
SELECT IF
SET
SORT CASES
STRING
SUBTITLE
TEMPORARY
TITLE
UPDATE
V2C
VALIDATEDATA
VALUE LABELS
VARIABLE ATTRIBUTE
VARSTOCASES
VECTOR

Nodo Modello Statistics

Il nodo Modello Statistics consente di analizzare e operare con i dati eseguendo le procedure IBM SPSS Statistics che generano PMML. Il nugget del modello che si creano possono essere utilizzati normalmente all'interno dei flussi IBM SPSS Modeler per il calcolo del punteggio e altre operazioni.

Per utilizzare questo nodo è necessario avere installata sul computer una versione compatibile di IBM SPSS Statistics con regolare licenza. Per ulteriori informazioni, consultare l'argomento "IBM SPSS Statistics Applicazioni di supporto" a pagina 312. Per informazioni relative alla compatibilità, consultare le note sul rilascio.

Le procedure analitiche IBM SPSS Statistics disponibili dipendono dal tipo di licenza che possiede l'utente.

Scheda Modello del nodo Modello Statistics

Nome modello. È possibile generare il nome del modello automaticamente in base al campo ID o obiettivo (oppure il tipo di modello nei casi non sia specificato tale campo) oppure indicare un nome personalizzato.

Seleziona una finestra di dialogo. Fare clic per visualizzare un elenco delle procedure IBM SPSS Statistics disponibili che è possibile selezionare ed eseguire. L'elenco include soltanto le procedure che generano PMML e per le quali si dispone di licenza e non include procedure scritte dall'utente.

1. Fare clic sulla procedura richiesta; viene visualizzata la finestra di dialogo IBM SPSS Statistics relativa.
2. Nella finestra di dialogo IBM SPSS Statistics, inserire i dettagli della procedura.
3. Fare clic su **OK** per tornare al nodo Modello Statistics; la sintassi di IBM SPSS Statistics viene visualizzata nella scheda Modello.
4. Per tornare alla finestra di dialogo IBM SPSS Statistics in qualsiasi momento, per esempio per modificare la propria query, fare clic sul pulsante di visualizzazione della finestra di dialogo IBM SPSS Statistics alla destra del pulsante di selezione della procedura.

Riepilogo del nugget del modello del nodo Modello Statistics

Quando si esegue il nodo Modello Statistics, questo esegue la procedura IBM SPSS Statistics associata e crea un nugget del modello che è possibile utilizzare nei flussi IBM SPSS Modeler per il calcolo del punteggio.

La scheda Riepilogo del nugget del modello visualizza informazioni sui campi, le impostazioni di creazione e l'elaborazione della stima del modello. I risultati vengono rappresentati in una vista della struttura ad albero che può essere espansa o compressa facendo clic su elementi specifici.

Il pulsante **Visualizza modello** visualizza i risultati in una forma modificata di IBM SPSS Statistics Output Viewer. Per ulteriori informazioni sul visualizzatore, vedere la documentazione relativa a IBM SPSS Statistics.

Dal menu File sono disponibili le normali opzioni di esportazione e stampa. Per ulteriori informazioni, consultare l'argomento "Visualizzazione dell'output" a pagina 272.

Nodo Output Statistics

Il nodo Output Statistics Output consente di richiamare una procedura IBM SPSS Statistics per analizzare i propri dati IBM SPSS Modeler. I risultati possono essere visualizzati in una finestra del browser oppure salvati nel formato del file di output IBM SPSS Statistics. Da IBM SPSS Modeler è possibile accedere a una vasta gamma di procedure analitiche di IBM SPSS Statistics.

Per utilizzare questo nodo è necessario avere installata sul computer una versione compatibile di IBM SPSS Statistics con regolare licenza. Per ulteriori informazioni, consultare l'argomento "IBM SPSS Statistics Applicazioni di supporto" a pagina 312. Per informazioni relative alla compatibilità, consultare le note sul rilascio.

Se necessario, è possibile visualizzare la scheda Filtro per filtrare o rinominare i campi in modo che risultino conformi agli standard di denominazione IBM SPSS Statistics. Per ulteriori informazioni, consultare l'argomento "Ridenominazione o filtro dei campi per IBM SPSS Statistics" a pagina 350.

Syntax Reference. Per informazioni dettagliate su specifiche procedure IBM SPSS Statistics, consultare la guida *IBM SPSS Statistics Command Syntax Reference*, inclusa nella copia del software di IBM SPSS Statistics. Per visualizzare la guida dalla scheda Sintassi, selezionare l'opzione **Editor di sintassi** e fare clic sul pulsante Avvia Guida sintassi IBM SPSS Statistics.

Scheda Sintassi del nodo Output Statistics

Utilizzare questa scheda per creare la sintassi per la procedura IBM SPSS Statistics che si desidera utilizzare per analizzare i dati. La sintassi è composta da due parti: un'istruzione e le opzioni associate a tale istruzione. L'istruzione specifica l'analisi o l'operazione da eseguire e i campi da utilizzare. Le opzioni specificano tutto il resto, comprese le statistiche da visualizzare, i campi derivati da salvare e così via.

Opzione finestra di dialogo IBM SPSS Statistics

Se non si conosce la sintassi di IBM SPSS Statistics per una procedura, il modo più semplice per creare una sintassi in IBM SPSS Modeler consiste nello scegliere l'opzione **Finestra di dialogo IBM SPSS Statistics**, selezionare la finestra di dialogo per la procedura, compilare la finestra e fare clic su OK. La sintassi verrà inserita nella scheda Sintassi del nodo IBM SPSS Statistics che si sta utilizzando in IBM SPSS Modeler. Sarà quindi possibile eseguire il flusso per ottenere l'output dalla procedura.

Se lo si desidera, è possibile generare un nodo origine File Statistics per l'importazione dei dati risultanti. Questo è utile, per esempio, se una procedura scrive campi, quali i punteggi, nell'insieme di dati attivo oltre a visualizzare l'output.

Per creare la sintassi:

1. Fare clic sul pulsante **Seleziona una finestra di dialogo**.
2. Scegliere una delle opzioni:
 - **Analizza.** Visualizza il contenuto del menu Analizza di IBM SPSS Statistics; scegliere la procedura che si desidera utilizzare.
 - **Altro.** Se disponibile, visualizza le finestre di dialogo create mediante il generatore di finestre di dialogo personalizzate in IBM SPSS Statistics e le altre finestre di IBM SPSS Statistics non visualizzate nel menu Analizza e per le quali si dispone di una licenza. In assenza di finestre di dialogo, l'opzione non viene visualizzata.

Nota: le finestre di dialogo Preparazione automatica dati non vengono visualizzate.

Se si dispone di una finestra di dialogo personalizzata di IBM SPSS Statistics che crea nuovi campi, tali campi non possono essere utilizzati in IBM SPSS Modeler perché il nodo Output Statistics è un nodo terminale.

Se si desidera, selezionare la casella **Genera un nodo di importazione per i dati risultanti** per creare un nodo origine File Statistics che può essere utilizzato per importare i dati risultanti in un altro stream. Il nodo viene inserito nell'area su schermo con i dati contenuti nel file *.sav* specificato nel campo **File** (la posizione di default corrisponde alla directory di installazione di IBM SPSS Modeler).

Opzione Editor di sintassi

Per salvare la sintassi creata per una procedura utilizzata frequentemente, procedere come descritto di seguito:

1. Fare clic sul pulsante Comandi relativi ai file (il primo sulla barra degli strumenti).
2. Scegliere **Salva** o **Salva con nome** dal menu.
3. Salvare il file come file *.sps*.

Per utilizzare i file di sintassi creati in precedenza dell'Editor di sintassi, sostituendo il contenuto corrente, se presente, procedere come descritto di seguito:

1. Fare clic sul pulsante Comandi relativi ai file (il primo sulla barra degli strumenti).
2. Scegliere **Apri** dal menu.
3. Selezionare un file *.sps* e incollarne il contenuto nella scheda Sintassi del nodo Output.

Per inserire la sintassi salvata in precedenza senza sostituire il contenuto corrente, procedere come descritto di seguito:

1. Fare clic sul pulsante Comandi relativi ai file (il primo sulla barra degli strumenti).
2. Scegliere **Inserisci** dal menu
3. Selezionare un file *.sps* e incollarne il contenuto nel punto del nodo Output specificato dal cursore.

Se si desidera, selezionare la casella **Genera un nodo di importazione per i dati risultanti** per creare un nodo origine File Statistics che può essere utilizzato per importare i dati risultanti in un altro stream. Il nodo viene inserito nell'area su schermo con i dati contenuti nel file *.sav* specificato nel campo **File** (la posizione di default corrisponde alla directory di installazione di IBM SPSS Modeler).

Quando si fa clic su **Esegui**, i risultati vengono visualizzati nel Visualizzatore output IBM SPSS Statistics. Per ulteriori informazioni sul visualizzatore, vedere la documentazione relativa a IBM SPSS Statistics.

Scheda Output del nodo Output Statistics

La scheda Output consente di specificare il formato e la posizione dell'output. Si può scegliere di visualizzare i risultati sullo schermo o di inviarli a uno dei tipi di file disponibili.

Nome output. Specifica il nome dell'output generato durante l'esecuzione del nodo. **Automatico** sceglie un nome in base al nodo che genera l'output. Se necessario, è possibile selezionare **Personalizzato** per specificare un nome diverso.

Output su schermo (default). Crea un nuovo oggetto di output da visualizzare in linea. L'oggetto di output verrà visualizzato nella scheda Output della finestra dei manager dopo l'esecuzione del nodo di output.

Output su file. Salva l'output in un file quando si esegue il nodo. Se si sceglie questa opzione, immettere un nome di file nel campo **Nome file** (oppure passare a una directory e specificare un nome di file utilizzando il pulsante di selezione dei file) e selezionare un tipo di file.

Tipo di file. Scegliere il tipo di file al quale si desidera inviare l'output.

- **Documento HTML (*.html).** Scrive l'output in formato HTML.
- **File di IBM SPSS Statistics Viewer (*.spv).** Scrive l'output in un formato che può essere letto da IBM SPSS Statistics Output Viewer.
- **File di IBM SPSS Statistics Web Reports (*.spw).** Scrive l'output in formato IBM SPSS Statistics Web Reports che può essere pubblicato in un repository IBM SPSS Collaboration and Deployment Services e successivamente visualizzato in un browser Web. Per ulteriori informazioni, consultare l'argomento "Pubblica sul Web" a pagina 273.

Nota: se si seleziona **Output su schermo** la direttiva IBM SPSS Statistics OMS VIEWER=NO non ha alcun effetto; inoltre, le API di script (*Di base* e modulo *Python SpssClient*) non sono disponibili in IBM SPSS Modeler.

Nodo Esporta Statistics

Il nodo Esporta Statistics consente di esportare i dati in formato *.sav* di IBM SPSS Statistics. I file IBM SPSS Statistics *.sav* possono essere letti da IBM SPSS Statistics Base e altri moduli. Questo formato viene inoltre utilizzato per i file cache di IBM SPSS Modeler.

In alcuni casi, il mapping dei nomi di campi IBM SPSS Modeler ai nomi di variabili IBM SPSS Statistics può causare errori perché i nomi di variabili IBM SPSS Statistics sono limitati a 64 caratteri e non possono contenere alcuni caratteri, come spazi, simboli dollaro (\$) e trattini (-). Esistono due modi per risolvere il problema:

- È possibile rinominare i campi in modo da soddisfare i requisiti per i nomi di variabili IBM SPSS Statistics facendo clic sulla scheda Filtro. Per ulteriori informazioni, consultare l'argomento "Ridenominazione o filtro dei campi per IBM SPSS Statistics" a pagina 350.
- Scegliere di esportare sia i nomi di campo che le etichette da IBM SPSS Modeler.

Nota: IBM SPSS Modeler scrive i file *.sav* in formato Unicode UTF-8. IBM SPSS Statistics supporta solo file in formato Unicode UTF-8 a partire dalla release 16.0. Per prevenire il rischio di corruzione dei dati, i file *.sav* salvati con codifica Unicode non devono essere utilizzati nelle versioni IBM SPSS Statistics precedenti alla 16.0. Per ulteriori informazioni, vedere l' Aiuto di IBM SPSS Statistics.

Insiemi di risposte multiple. Tutti gli insiemi di risposte multiple definiti nel flusso saranno automaticamente conservati quando il file viene esportato. Gli insiemi di risposte multiple si possono visualizzare e modificare da qualsiasi nodo mediante la scheda Filtro. Per ulteriori informazioni, consultare l'argomento "Modifica di insiemi di risposte multiple" a pagina 124.

Scheda Esporta del nodo Esporta Statistics

File di esportazione. Specificare il nome del file. Digitare un nome di file o fare clic sul pulsante di selezione dei file per individuare la posizione del file.

Tipo di file. Selezionare se il file deve essere salvato in formato normale *.sav* o compresso *.zsav*.

Crittografa file con password. Per proteggere il file con una password, selezionare questa casella; viene richiesto di immettere e confermare la **Password** in una finestra di dialogo separata.

Nota: I file protetti da password possono essere aperti solo da IBM SPSS Modeler versione 16 o successiva o da IBM SPSS Statistics versione 21 o successiva.

Esporta nomi di campo come variabile. Specifica un metodo di gestione di nomi ed etichette delle variabili durante l'esportazione da IBM SPSS Modeler in un file IBM SPSS Statistics *.sav* o *.zsav*.

- **Nomi ed etichette di variabile.** Selezionare questa opzione per esportare sia i nomi che le etichette dei campi di IBM SPSS Modeler. I nomi vengono esportati come nomi di variabili IBM SPSS Statistics, mentre le etichette vengono esportate come etichette di variabile IBM SPSS Statistics.
- **Nomi come etichette di variabile.** Selezionare questa opzione per utilizzare i nomi di campi IBM SPSS Modeler come etichette di variabile in IBM SPSS Statistics. IBM SPSS Modeler consente di includere nei nomi dei campi caratteri non validi nei nomi di variabili IBM SPSS Statistics. Per evitare di creare nomi IBM SPSS Statistics non validi, selezionare **etichette** oppure utilizzare la scheda Filtro per modificare i nomi di campo.

Avvia applicazione. Se installato nel computer, IBM SPSS Statistics può essere richiamato direttamente nel file di dati salvato tramite questa opzione. Le opzioni per il lancio dell'applicazione devono essere

specificate nella finestra di dialogo Applicazioni di supporto. Per ulteriori informazioni, consultare l'argomento "IBM SPSS Statistics Applicazioni di supporto" a pagina 312. Per creare semplicemente un file IBM SPSS Statistics *.sav* o *.zsav* senza aprire un programma esterno, deselezionare questa opzione.

Genera un nodo di importazione per questi dati. Selezionare questa opzione per generare automaticamente un nodo origine File Statistics che leggerà il file di dati esportato. Per ulteriori informazioni, consultare l'argomento "Nodo File Statistics" a pagina 342.

Ridenominazione o filtro dei campi per IBM SPSS Statistics

Prima di esportare o eseguire la distribuzione dei dati da IBM SPSS Modeler ad applicazioni esterne quali IBM SPSS Statistics, potrebbe essere necessario rinominare o correggere i nomi dei campi. Le finestre di dialogo Trasformazioni Statistics, Output Statistics ed Esporta Statistics contengono una scheda Filtro in grado di semplificare questo processo.

Una descrizione di base della funzione della scheda Filtro viene presentata in un'altra sezione. Per ulteriori informazioni, consultare l'argomento "Impostazione delle opzioni di filtro" a pagina 122. Questo argomento contiene suggerimenti per la lettura dei dati in IBM SPSS Statistics.

Per modificare i nomi dei campi in modo che corrispondano alle convenzioni di denominazione di IBM SPSS Statistics:

1. Nella scheda Filtro, fare clic sul pulsante della barra degli strumenti Comandi relativi al filtro (il primo sulla barra degli strumenti).
2. Selezionare Rinomina per IBM SPSS Statistics.
3. Nella finestra Rinomina per IBM SPSS Statistics, è possibile scegliere di sostituire i caratteri non validi nei nomi di file con un carattere **cancelletto (#)** o **di sottolineatura (_)**.

Rinomina insieme di risposte multiple. Selezionare questa opzione se si desidera modificare i nomi di insieme di risposte multiple, che possono essere importati in IBM SPSS Modeler utilizzando un nodo origine File Statistics. Vengono utilizzati per registrare i dati che possono avere più di un valore per ogni caso, ad esempio nelle risposte alle indagini.

Capitolo 9. Supernodi

Panoramica sui Supernodi

Uno dei motivi per cui l'interfaccia di programmazione visuale di IBM SPSS Modeler risulta semplice da utilizzare è che ogni nodo include una funzione chiaramente definita. Tuttavia, per l'elaborazione complessa può essere necessaria una lunga sequenza di nodi ed è pertanto possibile che lo spazio nell'area venga occupato e che diventi quindi difficile seguire i diagrammi del flusso. Per evitare la riduzione dello spazio in stream lunghi e complessi sono disponibili due metodi:

- È possibile suddividere una sequenza di elaborazione in più flussi collegati tra di loro. Per esempio, il primo stream crea un file di dati utilizzato dal secondo stream come input. Il secondo crea un file di dati utilizzato dal terzo come input e così via. Per gestire questi stream collegati è possibile salvarli in un **progetto**. Un progetto consente di organizzare più flussi con i relativi output. Tuttavia, un file di progetto contiene solo un riferimento agli oggetti inclusi e non elimina il problema della gestione di più file di stream.
- Un'alternativa più semplice per la gestione di elaborazioni di stream complesse consiste nel creare un **Supernodo**.

I Supernodi raggruppano più nodi in un unico nodo contenente sezioni di un flusso di dati. Questo metodo assicura numerosi vantaggi al data miner:

- I flussi sono più ordinati e gestibili.
- I nodi possono essere combinati in un Supernodo specifico per un settore di business.
- I Supernodi possono essere esportati in librerie in modo da essere riutilizzati in più progetti di data mining.

Tipi di Supernodi

Nel flusso di dati i Supernodi sono rappresentati da un'icona a forma di stella. L'icona è contraddistinta da ombreggiature diverse che indicano la direzione del flusso del flusso rispetto al Supernodo.

Sono disponibili tre tipi di Supernodi:

- Supernodi origine
- Supernodi di elaborazione
- Supernodo finale

Supernodi origine

I Supernodi origine contengono una sorgente dati come i normali nodi origine e possono essere utilizzati nelle stesse posizioni consentite ai normali nodi origine. L'ombreggiatura nella metà sinistra di un Supernodo origine indica che tale Supernodo è "chiuso" a sinistra e che la direzione del flusso dei dati deve essere *in entrata* nel Supernodo.

I Supernodi origine includono un solo punto di connessione a destra, che indica il flusso di dati dal Supernodo verso il flusso.

Supernodi di elaborazione

I Supernodi di elaborazione contengono solo nodi di elaborazione e non sono contrassegnati da ombreggiature in quanto il flusso dei dati è consentito sia *verso* sia *da* questo tipo di Supernodo.

I Supernodi di elaborazione includono punti di connessione sia a destra sia a sinistra, che indicano che i dati entrano nel Supernodo e quindi escono per tornare al flusso. Sebbene i Supernodi possano contenere frammenti di flussi aggiuntivi e ulteriori flussi, i punti di connessione devono seguire lo stesso percorso che connette il punto *Da flusso* al punto *A stream*.

Nota: i Supernodi di elaborazione sono a volte detti anche *Supernodi di manipolazione*.

Supernodo finale

I Supernodi terminali contengono uno o più nodi terminali (plot, tabelle e così via) e possono essere utilizzati nello stesso modo in cui vengono utilizzati i nodi terminali. L'ombreggiatura nella metà destra del Supernodo terminale indica che il Supernodo è "chiuso" a destra e che l'unica direzione consentita per il flusso dei dati è *in entrata* nel Supernodo terminale.

I Supernodi terminali contengono un solo punto di connessione a sinistra, che indica che il flusso dei dati entra nel Supernodo dal flusso e viene interrotto all'interno del Supernodo.

I Supernodi terminali possono inoltre contenere script utilizzati per specificare l'ordine di esecuzione per tutti i nodi terminali all'interno del Supernodo. Per ulteriori informazioni, consultare l'argomento "Supernodi e script" a pagina 357.

Creazione di Supernodi

La creazione di un Supernodo consente di "contrarre" il flusso di dati tramite l'incapsulamento di più nodi in un unico nodo. Dopo aver creato o caricato un flusso nell'area, è possibile creare un Supernodo in molti modi.

Selezione multipla

La procedura più semplice per la creazione di un Supernodo consiste nel selezionare tutti i nodi che si desidera incapsulare:

1. Utilizzare il mouse per selezionare più nodi nell'area del flusso. Per selezionare un flusso o una sezione di un flusso è inoltre possibile fare clic sugli elementi desiderati tenendo premuto Maiusc.
Nota: i nodi selezionati devono appartenere a un flusso continuo o biforcuto. Non è possibile selezionare nodi che non siano adiacenti o connessi in qualche modo.
2. Tramite una di queste procedure incapsulare quindi i nodi selezionati:
 - Fare clic sull'icona del Supernodo (in forma di stella) nella barra degli strumenti.
 - Fare clic con il pulsante destro del mouse sul Supernodo e dal menu di scelta rapida scegliere:
Crea supernodo > Da selezione
 - Dal menu Supernodo scegliere:
Crea supernodo > Da selezione

Queste tre opzioni consentono di incapsulare i nodi in un Supernodo la cui icona è ombreggiata in modo da indicare il tipo di Supernodo (di origine, di elaborazione o terminale) in base ai contenuti.

Selezione singola

È inoltre possibile creare un Supernodo selezionando un unico nodo e utilizzando le opzioni di menu per definirne inizio e fine oppure incapsulando tutti gli elementi a valle del nodo selezionato.

1. Fare clic sul nodo di inizio dell'incapsulamento.
2. Dal menu Supernodo scegliere:
Crea supernodo > Da qui

I Supernodi possono inoltre essere creati in modo più interattivo selezionando il nodo iniziale e quello terminale della sezione del flusso che si desidera incapsulare:

1. Fare clic sul primo o sull'ultimo nodo che si desidera includere nel Supernodo.
2. Dal menu Supernodo scegliere:
Crea supernodo > Seleziona...
3. In alternativa, è possibile utilizzare le opzioni del menu di scelta rapida facendo clic con il pulsante destro del mouse sul nodo desiderato.
4. Il cursore assume la forma dell'icona di un Supernodo per indicare che è necessario selezionare un altro punto nel flusso. Spostare il cursore a monte o a valle del frammento del Supernodo selezionato e fare clic su un nodo. Questa operazione determina la sostituzione di tutti i nodi inclusi tra i due punti selezionati con l'icona del Supernodo.

Nota: i nodi selezionati devono appartenere a un flusso continuo o biforcuto. Non è possibile selezionare nodi che non siano adiacenti o connessi in qualche modo.

Nidificazione di Supernodi

I Supernodi possono essere nidificati all'interno di altri Supernodi. Ai Supernodi nidificati si applicano le stesse regole valide per ogni tipo di Supernodo (origine, di elaborazione e terminale). Per esempio, per un Supernodo di elaborazione contenente Supernodi nidificati è necessario che il flusso di dati nei due sensi sia assicurato attraverso tutti i Supernodi nidificati. Se uno dei Supernodi nidificati è terminale, il flusso di dati nella gerarchia di Supernodi non viene eseguito.

I Supernodi origine e terminali possono contenere altri tipi di Supernodi nidificati, ma anche in questo caso sono valide le stesse regole di base per la creazione di Supernodi.

Blocco di Supernodi

Una volta creato un Supernodo, è possibile bloccarlo con una password per impedire che venga modificato. Questa protezione può risultare per esempio necessaria nella creazione di stream, o parti di stream, come modelli a valore fisso che devono essere utilizzati nell'organizzazione da addetti con meno esperienza nell'impostazione delle interrogazioni di IBM SPSS Modeler.

Il blocco di un Supernodo non impedisce agli utenti di immettere nella scheda Parametri valori relativi agli eventuali parametri definiti; un Supernodo bloccato può essere inoltre eseguito senza digitare la password.

Nota: non è possibile eseguire il blocco e lo sblocco utilizzando gli script.

Blocco e sblocco di Supernodi

Avvertenza: non è possibile recuperare le password perse.

È possibile bloccare o sbloccare un Supernodo da una delle tre schede.

1. Fare clic su **Blocca nodo**.
2. Digitare e confermare la password.
3. Fare clic su **OK**.

Un Supernodo protetto da password è riconoscibile nell'area del flusso dal simbolo del lucchetto che compare nella parte superiore sinistra dell'icona Supernodo.

Sblocco di un Supernodo

1. Per eliminare in modo definitivo la protezione della password, fare clic su **Sblocca nodo**; viene richiesto di digitare la password.

2. Digitare la password e fare clic su **OK**; il Supernodo non è più protetto da password e nel flusso scompare il simbolo del lucchetto accanto all'icona.

Modifica di un Supernodo bloccato

Se l'utente tenta di definire parametri o di effettuare lo zoom in avanti per visualizzare un Supernodo bloccato, viene richiesto di immettere la password.

Immettere la password e fare clic su **OK**.

L'utente è ora in grado di modificare le definizioni dei parametri ed eseguire lo zoom in avanti o indietro tutte le volte che lo desidera fino a quando non chiude il flusso in cui si trova il Supernodo.

Tenere presente che questo non elimina la protezione della password, ma consente solo l'accesso all'utente per l'utilizzo del Supernodo. Per ulteriori informazioni, consultare l'argomento "Blocco e sblocco di Supernodi" a pagina 353.

Modifica Supernodo

Dopo aver creato un Supernodo è possibile analizzarlo più da vicino ingrandendolo; se il Supernodo è bloccato, verrà richiesta l'immissione della password. Per ulteriori informazioni, consultare l'argomento "Modifica di un Supernodo bloccato".

Per visualizzare i contenuti di un Supernodo è possibile utilizzare l'icona della modalità Zoom avanti del Supernodo disponibile nella barra degli strumenti di IBM SPSS Modeler oppure il metodo seguente:

1. Fare clic con il pulsante destro del mouse su un Supernodo.
2. Dal menu di scelta rapida scegliere **Zoom avanti**.

I contenuti del Supernodo selezionato verranno visualizzati in un ambiente di IBM SPSS Modeler leggermente diverso, contenente i connettori che indicano il flusso di dati nel flusso o nel frammento di stream. A questo livello nell'area del flusso, è possibile eseguire diverse attività:

- Modificare il tipo di supernodo — origine, elaborazione o terminale.
- Creare parametri o modificare i valori di un parametro. I parametri vengono utilizzati negli script e nelle espressioni CLEM.
- Specificare opzioni di cache per il Supernodo e i relativi sottonodi.
- Creare o modificare lo script di un Supernodo (solo per i Supernodi terminali).

Modifica dei tipi di Supernodo

In alcuni casi può essere utile modificare il tipo di un Supernodo. Questa opzione è disponibile e si applica solo al livello della modalità Zoom avanti di un Supernodo. I tre tipi di Supernodo sono illustrati nella tabella riportata di seguito.

Tabella 57. Tipi di Supernodo.

Tipo di Supernodo	Descrizione
Supernodo origine	Una sola connessione in uscita
Supernodo di elaborazione	Due connessioni: una in entrata e una in uscita
Supernodo terminale	Una sola connessione in entrata

Modifica del tipo di un Supernodo

1. Assicurarsi di essere nella modalità Zoom avanti del Supernodo.
2. Dal menu Supernodo, selezionare **Tipo Supernodo** e quindi scegliere il tipo.

Annotazione e ridenominazione dei Supernodi

È possibile rinominare un Supernodo visualizzato nel flusso nonché scrivere annotazioni per progetti o report. Per accedere a queste proprietà:

- Fare clic con il pulsante destro del mouse su un Supernodo (nella modalità Zoom indietro del Supernodo) e quindi scegliere **Rinomina e annota**.
- In alternativa, scegliere **Rinomina e annota** dal menu Supernodo. Questa opzione è disponibile in entrambe le modalità Zoom avanti e Zoom indietro.

In entrambi i casi verrà aperta una finestra di dialogo con la scheda Annotazioni selezionata. Utilizzare le opzioni disponibili per personalizzare il nome visualizzato nell'area del flusso e fornire informazioni sulle operazioni eseguite per il Supernodo.

Utilizzo di commenti con i Supernodi

Se si crea un Supernodo da un nodo o da un nugget per il quale sono stati creati dei commenti, è necessario includere tali commenti nella selezione per la creazione del Supernodo se si desidera che i commenti appaiano nello stesso. SE non viene incluso nella selezione, il commento resta nel flusso quando viene creato il Supernodo.

Quando si espande un Supernodo che include commenti, i commenti vengono reintegrati nella posizione in cui si trovavano prima della creazione del Supernodo.

Quando si espande un Supernodo che include oggetti commentati senza che i commenti siano inclusi nel Supernodo, gli oggetti vengono reintegrati nella loro posizione originale senza che i commenti vengano nuovamente associati.

Parametri dei Supernodi

In IBM SPSS Modeler è possibile impostare variabili definite dall'utente, per esempio *Valminimo*, i cui valori possono essere specificati per l'utilizzo in script o espressioni CLEM. Queste variabili vengono denominate **parametri**. È possibile impostare parametri per stream, sessioni e Supernodi. Tutti i parametri impostati per un Supernodo sono disponibili durante la creazione di espressioni CLEM per il Supernodo specifico o qualsiasi nodo nidificato. Le impostazioni di parametri relativi a Supernodi nidificati non sono disponibili per il Supernodo padre.

Per la creazione e l'impostazione di parametri relativi a Supernodi è necessario eseguire due passaggi:

1. Definire i parametri per il Supernodo.
2. Specificare quindi il valore per ogni parametro del Supernodo.

Questi parametri possono essere utilizzati in espressioni CLEM per tutti i nodi incapsulati.

Definizione dei parametri di un Supernodo

È possibile definire i parametri per un Supernodo in entrambe le modalità Zoom avanti e Zoom indietro. I parametri definiti si applicano a tutti i nodi incapsulati. Per definire i parametri di un Supernodo è innanzitutto necessario accedere alla scheda Parametri della finestra di dialogo del Supernodo. Utilizzare uno dei metodi seguenti per aprire la finestra di dialogo:

- Fare doppio clic su un Supernodo nel flusso.
- Dal menu Supernodo scegliere **Imposta parametri**.
- In alternativa, nella modalità Zoom avanti di un Supernodo scegliere **Imposta parametri** dal menu di scelta rapida.

Dopo aver aperto la finestra di dialogo, la scheda Parametri è visibile per qualsiasi parametro definito in precedenza.

Per definire un nuovo parametro

Fare clic sul pulsante **Definisci parametri** per aprire la finestra di dialogo.

Nome. In questa colonna sono visualizzati i nomi dei parametri. È possibile creare un nuovo parametro inserendo nel campo un nome, Per esempio, per creare un parametro per la temperatura minima, è possibile digitare *valminimo*. Non includere il prefisso \$P- che indica un parametro nelle espressioni CLEM. Tale nome è inoltre quello visualizzato nel Generatore espressioni CLEM.

Nome lungo. Elenca il nome descrittivo di ogni parametro creato.

Archiviazione. Selezionare un tipo di archiviazione dall'elenco. La classe archiviazione indica come sono stati archiviati i valori dei dati nel parametro. Per esempio, quando si utilizzano valori contenenti zero iniziali che si desidera mantenere (per esempio 008), è necessario selezionare **Stringa** come tipo di archiviazione. In caso contrario, gli zeri verranno rimossi dal valore. I tipi di archiviazione disponibili sono stringa, intero, reale, ora, data e timestamp. Per i parametri data, tenere presente che è necessario specificare i valori con la notazione standard ISO, come illustrato di seguito.

Valore. Elenca il valore corrente di ogni parametro. Modificare il parametro in base alle esigenze specifiche. Si noti che per specificare i parametri data è necessario utilizzare la notazione standard ISO (ovvero, AAAA-MM-GG). Le date specificate in altri formati non sono accettate.

Tipo (facoltativo). Se si pianifica la distribuzione del flusso in un'applicazione esterna, selezionare un livello di misurazione dall'elenco. In caso contrario, si consiglia di non modificare la colonna *Tipo*. Se si desidera specificare dei vincoli per il valore del parametro quale un limite superiore e inferiore per un intervallo numerico, selezionare **Specificata** dall'elenco.

Si noti che è possibile impostare le opzioni nome lungo, classe archiviazione e tipo per i parametri solo tramite l'interfaccia utente e non tramite script.

Fare clic sulle frecce a destra per spostare il parametro selezionato più in alto o in basso nell'elenco dei parametri. Per rimuovere il parametro selezionato, utilizzare il pulsante di eliminazione contrassegnato da una X.

Impostazione dei parametri di un Supernodo

Dopo aver definito i parametri di un Supernodo, è possibile specificare i valori utilizzando i parametri in script o espressioni CLEM.

Per specificare i parametri di un Supernodo

1. Fare doppio clic sull'icona del Supernodo per aprire la finestra di dialogo del Supernodo.
2. In alternativa, scegliere **Imposta parametri** dal menu Supernodo.
3. Fare clic sulla scheda **Parametri**. *Nota:* i campi inclusi in questa finestra di dialogo sono i campi definiti facendo clic sul pulsante **Definisci parametri** in questa scheda.
4. Immettere un valore nella casella di testo per ogni parametro creato. Per esempio, è possibile impostare il valore *valminimo* su una particolare soglia significativa. Questo parametro può quindi essere utilizzato in numerose operazioni, quali la selezione a scopo di ulteriori esplorazioni dei record superiori o inferiori alla soglia specificata.

Utilizzo dei parametri di un Supernodo per l'accesso alle proprietà del nodo proprietà

I parametri del Supernodo possono inoltre essere utilizzati per definire le proprietà del nodo (denominate anche **parametri di slot**) per i nodi incapsulati. Per esempio, si supponga di voler specificare che un Supernodo esegua l'addestramento di un nodo Rete neurale incapsulato per un determinato periodo di tempo tramite un campione casuale dei dati disponibili. Se si utilizzano i parametri, sarà possibile specificare valori relativi al periodo di tempo e alla percentuale del campione.

Si supponga che il Supernodo di esempio contenga un nodo Campione denominato *Campione* ed un nodo Rete neurale denominato *Addestra*. È possibile utilizzare le finestre di dialogo del nodo per impostare **Campione** del nodo Campione su % **casuale** e **Condizione arresto** del nodo Rete neurale su **Ora**. Dopo aver specificato queste opzioni, sarà possibile accedere alle proprietà del nodo tramite i parametri e indicare valori specifici per il Supernodo. Nella finestra di dialogo Supernodo, fare clic su **Definisci parametri** e creare i parametri visualizzati nella seguente tabella.

Tabella 58. Parametri da creare

Parametro	Valore	Nome lungo
Addestra.time	5	Tempo di addestramento (minuti)
Campione.random	10	Percentuale campione casuale

Nota: i nomi dei parametri, come *Campione.random*, utilizzano la sintassi corretta per fare riferimento alle proprietà del nodo, dove *Campione* rappresenta il nome del nodo e *random* è una proprietà del nodo.

Dopo aver definito questi parametri, è possibile modificare valori per le proprietà dei nodi Campione e Rete neurale senza riaprire le relative finestre di dialogo. È infatti sufficiente scegliere **Imposta parametri** dal menu Supernodo per accedere alla scheda Parametri della finestra di dialogo del Supernodo, in cui è possibile specificare nuovi valori per % **casuale** e **Ora**. Questo risulta particolarmente utile quando si analizzano i dati durante numerose iterazioni di creazione del modello.

Supernodi e cache

Dall'interno di un Supernodo è possibile memorizzare nella cache tutti i nodi eccetto i nodi terminali. Per controllare l'esecuzione della cache, fare clic con il pulsante destro del mouse su un nodo e scegliere una delle numerose opzioni dal menu di scelta rapida Cache. Questa opzione di menu è disponibile sia all'interno sia all'esterno di un Supernodo, nonché per i nodi incapsulati in un Supernodo.

Per le cache dei Supernodi esistono diverse regole:

- Se per uno dei nodi incapsulati in un Supernodo è attiva la cache, anche per il Supernodo sarà attiva la cache.
- La disattivazione della cache di un Supernodo determina la disattivazione della cache per *tutti* i nodi incapsulati.
- L'attivazione della cache in un Supernodo determina in realtà l'attivazione della cache per l'ultimo sottonodo che supporta la cache. In altre parole, se l'ultimo sottonodo è un nodo Selezione, la cache verrà attivata per tale nodo. Se l'ultimo sottonodo è un nodo finale (che non consente l'esecuzione della cache) la cache verrà attivata per il primo nodo a monte in grado di supportarla.
- Dopo aver impostato cache per i sottonodi di un Supernodo, qualsiasi attività di upstream dal nodo memorizzato nella cache, per esempio l'aggiunta o la modifica di nodi, determinerà lo svuotamento delle cache.

Supernodi e script

È possibile utilizzare il linguaggio di script di IBM SPSS Modeler per scrivere programmi semplici che consentono di manipolare ed eseguire i contenuti di un Supernodo terminale. Per esempio, è possibile che si desideri specificare l'ordine di esecuzione per un flusso complesso. Se, per esempio, un Supernodo contiene un nodo Calcola globali che deve essere eseguito prima di un nodo Plot, è possibile creare uno script che esegua il nodo Calcola globali per primo. I valori calcolati da questo nodo, quali la media o la deviazione standard, possono quindi essere utilizzati quando viene eseguito il nodo Plot.

La scheda Script della finestra di dialogo del Supernodo è disponibile solo per i Supernodi terminali.

Per aprire la finestra di dialogo di script per un Supernodo terminale

- Fare clic con il pulsante destro del mouse nell'area del Supernodo e scegliere **Script Supernodo**.

- In alternativa, sia in modalità Zoom avanti che Zoom indietro è possibile scegliere **Script del Supernodo** dal menu Supernodo.

Nota: gli script dei Supernodi vengono eseguiti solo per stream e Supernodi se nella finestra di dialogo è stata selezionata l'opzione **Esegui questo script**.

Le opzioni specifiche per gli script ed il relativo utilizzo in IBM SPSS Modeler sono illustrate nella *Guida per script e automazione*, disponibile nel DVD di IBM SPSS Modeler.

Salvataggio e caricamento dei Supernodi

Uno dei vantaggi dei Supernodi è la possibilità di salvarli e riutilizzarli in altri flussi. Per il salvataggio e il caricamento di Supernodi, viene utilizzata un'estensione *slb*.

Per salvare un Supernodo

1. Utilizzare la modalità Zoom avanti per il Supernodo.
2. Dal menu Supernodo scegliere **Salva Supernodo**.
3. Specificare un nome di file e una directory nella finestra di dialogo.
4. Specificare se si desidera aggiungere il Supernodo salvato al progetto corrente.
5. Fare clic su **Salva**.

Per caricare un Supernodo

1. Dal menu Inserisci nella finestra di IBM SPSS Modeler, scegliere **Supernodo**.
2. Selezionare un file di Supernodo (*slb*) dalla directory corrente o passare a una directory diversa.
3. Fare clic su **Carica**.

Nota: nei Supernodi importati, tutti i parametri hanno i valori di default. Per modificare i parametri, fare doppio clic su un Supernodo nell'area del flusso.

Note

Queste informazioni sono state preparate per prodotti e servizi offerti in tutto il mondo.

È possibile che IBM non offra i prodotti, servizi o funzioni illustrati in questa documentazione. Consultare il rappresentante locale IBM per le informazioni sui prodotti e servizi attualmente disponibili nella propria zona. Qualsiasi riferimento a un prodotto, programma o servizio IBM non implica o intende dichiarare che può essere utilizzato solo quel prodotto, programma o servizio IBM. In sostituzione a quelli forniti da IBM è possibile utilizzare qualsiasi prodotto, programma o servizio funzionalmente equivalente che non comporti la violazione dei diritti di proprietà intellettuale IBM o altri diritti. Tuttavia, è responsabilità dell'utente valutare e verificare il funzionamento di qualsiasi prodotto, programma o servizio non IBM.

IBM potrebbe avere brevetti o domande di brevetti in corso relativi ad argomenti discussi nella presente pubblicazione. La consegna del presente documento non conferisce alcuna licenza rispetto a questi brevetti. Rivolgere per iscritto i quesiti sulle licenze a:

IBM Director of Licensing
IBM Europe
Schoenaicher Str.220
D-7030 Boeblingen
Deutschland

Per richieste di licenze relative ad informazioni double-byte (DBCS) contattare il Dipartimento di Proprietà Intellettuale IBM nel proprio paese o inviare richieste per iscritto a:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
1623-14, Shimotsuruma, Yamato-shi
Kanagawa 242-8502 Japan

Il seguente paragrafo non è valido nel Regno Unito o per tutti i paesi le cui leggi nazionali siano in contrasto con le disposizioni in esso contenute INTERNATIONAL BUSINESS MACHINES CORPORATION FORNISCE QUESTA PUBBLICAZIONE "NELLO STATO IN CUI ESSA SI TROVA" SENZA ALCUNA GARANZIA ESPLICITA O IMPLICITA IVI INCLUSE EVENTUALI GARANZIE DI COMMERCIALIZZABILITÀ ED IDONEITÀ AD UNO SCOPO PARTICOLARE Alcuni stati non consentono limitazioni di garanzie espresse o implicite in determinate transazioni, pertanto quanto sopra potrebbe non essere applicabile.

Le presenti informazioni possono includere imprecisioni tecniche o errori tipografici. Le modifiche periodiche apportate alle informazioni contenute in questa pubblicazione verranno inserite nelle nuove edizioni della pubblicazione. IBM si riserva il diritto di apportare miglioramenti e/o modifiche al prodotto o programma descritto in questa pubblicazione in qualsiasi momento e senza preavviso.

Qualsiasi riferimento nelle presenti informazioni a siti Web non IBM viene fornito esclusivamente per facilitare la consultazione e non rappresenta in alcun modo un'approvazione o sostegno da parte nostra di tali siti Web. I materiali disponibili sui siti Web non fanno parte di questo prodotto IBM e l'utilizzo di questi è a discrezione dell'utente.

IBM può utilizzare o distribuire qualsiasi informazione fornita dall'utente nel modo che ritiene più idoneo senza incorrere in alcun obbligo nei confronti dell'utente stesso.

Coloro che detengono la licenza su questo programma e desiderano avere informazioni su di esso allo scopo di consentire: (i) lo scambio di informazioni tra programmi indipendenti ed altri (compreso questo) e (ii) l'uso reciproco di tali informazioni dovrebbero contattare:

IBM Software Group
ATTN: Licensing
200 W. Madison St.
Chicago, IL; 60606
U.S.A.

Tali informazioni saranno fornite in conformità ai termini e alle condizioni in vigore e, in alcuni casi, dietro pagamento.

Il programma su licenza descritto in questa documentazione e tutto il materiale su licenza ad esso relativo vengono forniti da IBM nei termini del Customer Agreement IBM IBM International Program License Agreement o di eventuali accordi equivalenti intercorsi tra le parti.

Tutti i dati sulle prestazioni qui contenuti sono stati elaborati in ambiente controllato. Di conseguenza, i risultati ottenuti con sistemi operativi diversi possono variare in modo significativo. Alcune misurazioni potrebbero essere state effettuate su sistemi in corso di sviluppo e non c'è garanzia che tali misurazioni coincidano con quelle effettuate sui sistemi comunemente disponibili. Inoltre, alcune misurazioni potrebbero essere stime elaborate tramite l'estrapolazione. I risultati effettivi potrebbero variare. Gli utenti di questo documento devono verificare i dati relativi al proprio ambiente specifico.

le informazioni relative a prodotti non IBM sono state ottenute dai fornitori di tali prodotti, da loro annunci pubblicati e da altre fonti disponibili al pubblico. IBM non ha testato quei prodotti e non può garantire l'accuratezza delle prestazioni, la compatibilità o qualsiasi altra dichiarazione relativa a prodotti non-IBM. Eventuali domande in merito alle funzionalità dei prodotti non IBM vanno indirizzate ai fornitori di tali prodotti.

Qualsiasi affermazione relativa agli obiettivi e alla direzione futura di IBM è soggetta a modifica o revoca senza preavviso e concerne esclusivamente gli scopi dell'azienda.

Le presenti informazioni includono esempi di dati e report utilizzati in operazioni di business quotidiane. Per fornire una descrizione il più possibile esaustiva, gli esempi includono nomi di persone, società, marchi e prodotti. Tutti questi nomi sono fittizi e ogni somiglianza a nomi e indirizzi utilizzati da aziende reali è puramente casuale.

Per chi visualizza queste informazioni a video: le fotografie e le illustrazioni a colori potrebbero non essere disponibili.

Marchi

IBM, il logo IBM e ibm.com sono marchi o marchi registrati di International Business Machines Corp., registrati in molte giurisdizioni nel mondo. Altri nomi di prodotti e servizi possono essere marchi di IBM o altre società. Un elenco aggiornato di marchi IBM è disponibile sul sito Web "Copyright and trademark information" all'indirizzo www.ibm.com/legal/copytrade.shtml.

Intel, il logo Intel, Intel Inside, il logo Intel Inside, Intel Centrino, il logo Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium e Pentium sono marchi o marchi registrati di Intel Corporation o delle sue consociate negli Stati Uniti e in altri paesi.

Linux è un marchio registrato di Linus Torvalds negli Stati Uniti e/o negli altri paesi.

Microsoft, Windows, Windows NT e il logo Windows sono marchi di Microsoft Corporation negli Stati Uniti e/o negli altri paesi.

UNIX è un marchio registrato di The Open Group negli Stati Uniti e in altri paesi.

Java e tutti i marchi e logo basati su Java sono marchi o marchi registrati di Oracle e/o suoi affiliati.

Altri nomi di prodotti e servizi possono essere marchi commerciali di IBM o di altre aziende.

Glossario

C

Covarianza. Una misura non standardizzata di associazione tra due variabili, pari alla deviazione del prodotto incrociato divisa per $N-1$.

K

Curtosi. Una misura di quanto le osservazioni si trovino raggruppate in un punto centrale. Per la distribuzione normale, il valore della statistica di curtosi è zero. Una curtosi positiva indica che, rispetto a una distribuzione normale, le osservazioni sono più raggruppate intorno al centro della distribuzione e hanno code più sottili fino ai valori estremi della distribuzione; a quel punto, le code della distribuzione leptocurtica sono più spesse rispetto a una distribuzione normale. Una curtosi negativa indica che, rispetto a una distribuzione normale, le osservazioni sono meno raggruppate e hanno code più spesse fino ai valori estremi della distribuzione; a quel punto, le code della distribuzione platicurtica sono più sottili rispetto a una distribuzione normale.

M

Massimo. Il valore più alto di una variabile numerica.

Media. Una misura di tendenza centrale. Media aritmetica, ovvero somma divisa per il numero di casi.

Median. È il valore sopra il quale e sotto il quale ricade la metà dei casi, il 50-esimo percentile. Se il numero di casi è pari, la mediana è pari alla media dei due casi centrali quando questi sono ordinati secondo l'ordine ascendente o discendente. La mediana è una misura di tendenza centrale non sensibile ai valori anomali, a differenza della media che può essere influenzata da valori eccezionalmente bassi o alti.

Minimo. Il valore più basso di una variabile numerica.

Modalità. Il valore che ricorre più frequentemente. Se più valori condividono la maggiore ricorrenza, ognuno di essi è una modalità.

R

Intervallo. La differenza tra il valore massimo ed il valore minimo di una variabile numerica.

S

Asimmetria. Una misura dell'asimmetria di una distribuzione. La distribuzione normale è simmetrica e ha un valore di asimmetria pari a 0. Una distribuzione con una notevole asimmetria positiva ha una lunga coda a destra. Una distribuzione con asimmetria negativa ha una coda a sinistra. In generale un'asimmetria con valore più che doppio dell'errore standard indica lo scostamento dalla normale simmetria.

Deviazione standard. Una misura di dispersione intorno alla media, uguale alla radice quadrata della varianza. L'unità di misura della deviazione standard è la stessa della variabile originale.

Deviazione standard. Una misura di dispersione intorno alla media. In una distribuzione normale, il 68% dei casi rientra in una deviazione standard della media e il 95% dei casi rientra in due deviazioni standard. Se, ad esempio, in una popolazione con distribuzione normale l'età media fosse 45 e la deviazione standard 10, il 95% dei casi cadrebbe fra 25 e 65 anni.

Errore standard. Una misura di quanto il valore di una statistica del test può variare da campione a campione. È la deviazione standard della distribuzione di campionamento di una statistica. L'errore standard della media, ad esempio, è la deviazione standard delle medie del campione.

Errore standard della curtosi. Il rapporto fra la curtosi di una distribuzione e il suo errore standard viene usato come test di normalità. L'ipotesi di normalità può essere rifiutata se questo rapporto è inferiore di 2 o maggiore di +2. Un valore positivo elevato per la curtosi indica che le code della distribuzione sono più lunghe di quelle di una distribuzione normale; un valore negativo per la curtosi indica code più corte, simili a quelle di una distribuzione uniforme a forma di scatola.

Errore standard della media. Una misura di quanto il valore della media può variare tra campioni presi dalla stessa distribuzione. Può essere utilizzata per confrontare genericamente la media osservata rispetto a un valore ipotizzato (ovvero, è possibile concludere che i due valori sono diversi se il rapporto della differenza rispetto all'errore standard è inferiore a -2 o maggiore di +2).

Errore standard dell'asimmetria. Il rapporto fra la simmetria di una distribuzione e il suo errore standard viene usato come test di normalità. L'ipotesi di normalità può essere rifiutata se questo rapporto è inferiore di 2 o maggiore di +2. Un valore positivo elevato per l'asimmetria indica una coda a destra lunga; un valore negativo estremo indica una coda a sinistra lunga.

Somma. La somma o il totale di tutti i valori non mancanti di tutti i casi.

U

Univocità. Valuta tutti gli effetti simultaneamente, correggendo ogni effetto per tutti gli altri effetti di qualunque tipo.

R

Validi. I casi validi il cui valore non è né il valore mancante di sistema né un valore definito come mancante dall'utente.

Varianza. Una misura della dispersione dei valori intorno alla media. È calcolata come somma dei quadrati degli scostamenti dalla media, divisa per un valore corrispondente alla somma dei casi meno uno. La varianza è espressa in quadrati dell'unità di misura della variabile.

Indice analitico

A

- accorpamento dei dati di serie temporali 153
- aggiunta
 - record 69
- aggregazione dei dati di serie temporali 153
- aggregazione di record 147
- alterna 254, 263
- animazione nei grafici 166
- anno fiscale
 - nodo Intervalli di tempo 157
- anonimizzazione dei nomi di campi 124
- ANOVA
 - nodo Medie 296
- ANOVA monodirezionale
 - nodo Medie 296
- anti-unione 74
- apertura
 - oggetti di output 272
- applicazioni di supporto 312
- archiviazione 116
 - conversione 132, 133
- archiviazione di campi
 - conversione 132
- aree nei grafici 248
- assegnazione di tipi di dati 95
- associazione per colonna 323
- associazione per riga 323
- attributi
 - nelle mappe 198
- attributi del campo 120
- attributi del tipo 120
- attualità
 - impostazione della data relativa 72
- autoregressione
 - modelli ARIMA 89

B

- bacchetta magica nei grafici 250
- bin decile 140
- bin percentile 140
- bin quartile 140
- bin quintile 140
- bin ventile 140
- blocchi
 - nodo Raccolta 140
- blocco di Supernodi 353
- browser del nodo Analisi
 - interpretazione 282
- browser del nodo Esplora
 - generazione di grafici 290
 - generazione di nodi 290
 - Menu File 285
 - Menu Modifica 285
- browser del nodo Qualità
 - generazione di nodi Filtro 289
 - generazione di nodi Seleziona 290
- browser del nodo Report 300

- browser del nodo Statistiche
 - generazione di nodi Filtro 295
 - interpretazione 295
 - Menu Genera 295
- browser del nodo Tabella
 - Menu Genera 277
 - ricerca 277
 - riordinamento delle colonne 275, 277
 - selezione di celle 275, 277

C

- cache
 - Supernodi 357
- calcola durate
 - preparazione automatica dati 99
- calcolo del punteggio
 - opzioni relative al nodo Valutazione 240
- calcolo durate
 - preparazione automatica dati 99
- campi
 - anonimizzazione dei dati 133
 - derivazione di più campi 128
 - etichette di campi e valori 116
 - riordinamento 163
 - selezione multipla 129
 - trasposizione 150
- campi chiave 69, 147
- campi chiavi primarie
 - Nodo di esportazione del database 318
- campi delle etichette
 - assegnazione di etichette ai record in output 119
- campi partizione 119, 146
- campionamento 1-ogni-n 64
- campionamento di dati contigui 64
- campioni di addestramento
 - bilanciamento 68
 - partizionamento dei dati 146
- campioni di convalida
 - partizionamento dei dati 146
- campioni di test
 - partizionamento dei dati 146
- campioni non casuali 63, 64
- campioni ponderati 66
- campioni raggruppati 63, 64, 66
- campioni sistematici 63, 64
- campioni stratificati 63, 64, 66, 67
- campo Conteggio
 - nodo Intervalli di tempo 153
 - padding o aggregazione di serie temporali 153
- campo EtichettaTempo
 - nodo Intervalli di tempo 152
- campo IndiceTempo
 - nodo Intervalli di tempo 152
- cancellare valori 59
- caratteri di commento
 - in nodi Testo variabile 20

- Caratteri EOL 20
- caratteristiche
 - nelle mappe 198
- caricamento di massa 323, 324
- categorizzazione con supervisione 142
- chi-quadrato
 - nodo Matrice 279
- Chi-quadrato di Pearson
 - nodo Matrice 279
- chiavi contigue 69
- cluster 254, 263
- codifica fittizia 147
- Cognos, vedere IBM Cognos BI 37
- collegamenti
 - nodo Web 225
- comando CREATE INDEX 321
- combinazione di dati 80
 - da più file 74
- commenti
 - utilizzo con i Supernodi 355
- concatenazione di record 80
- condizioni
 - definizione di una serie 131
 - specificata per un'unione 77
- connessioni
 - IBM SPSS Collaboration and Deployment Services Repository 8
- connessioni database
 - definizione 15
 - valori preimpostati 15
- conteggi
 - nodo Raccolta 140
 - output del nodo Statistiche 295
- conteggio uguale
 - nodo Raccolta 140
- contrassegno di elementi 248, 250
- controllo dei tipi 119
- conversione dei livelli di misurazione 114
- conversione di insiemi in flag 147, 148
- coordinate polari 262
- copia degli attributi del tipo 120
- copia delle visualizzazioni 266
- coropleta 176
 - esempio 192
- correlazioni 294
 - etichette descrittive 294
 - in valore assoluto 294
 - nodo Medie 298
 - output del nodo Statistiche 295
 - probabilità 294
 - significatività 294
- correlazioni di Pearson
 - nodo Medie 298
 - output del nodo Statistiche 295
- costi
 - grafici di valutazione 238
- creazione
 - nuovi campi 127, 128
- creazione di flag 147, 149
- creazione istanza 112, 114, 115

CRISP-DM
Data Understanding 7

D

data/ora 112
database
 caricamento di massa 323, 324
database ADO
 importazione 26
database di grandi dimensioni 61
 esecuzione di un'esplorazione dei
 dati 284
database In2data
 importazione 26
database Quanvert
 importazione 26
dataset principale 80
date
 impostazione di formati 121
dati
 aggregazione 69
 analisi 61
 anonimizzazione 133
 archiviazione 132, 133
 esplorazione 284
 preparazione 61
 tipo di archiviazione 116
dati annuali
 nodo Intervalli di tempo 156
dati caso
 nodo di origine IBM SPSS Data
 Collection 25, 26
dati categoriali 114
dati continui 114, 117
dati CSV
 importazione 26
dati di campionamento 67
dati di indagine
 importazione 26, 29
 nodo di origine IBM SPSS Data
 Collection 25
dati di indagine IBM SPSS Data
Collection
 importazione 25, 26
dati di riepilogo 69
dati di serie temporali
 aggregazione 151, 153
 assegnazione di etichette 151, 152,
 153, 154
 creazione dai dati 153
 definizione 151, 152, 153, 154
 intervalli 152
 padding 151, 153
 periodo di stima 154
 record di holdout 154
dati di testo a campi fissi 22
dati di testo a campi liberi 20
dati di testo delimitati 20
dati distorti 68
dati mensili
 nodo Intervalli di tempo 158
dati nominali 118
dati non distorti 68
dati ordinali 118
dati Quancept
 importazione 26
dati Quantum
 importazione 26
dati ricerche di mercato
 importazione 26, 29
 nodo di origine IBM SPSS Data
 Collection 25, 29
dati ritardati 162
dati sbilanciati 68
dati settimanali
 nodo Intervalli di tempo 158
dati simulati
 nodo Genera simulazione 42
dati sintetici
 nodo Input utente 55
dati Surveycraft
 importazione 26
dati trimestrali
 nodo Intervalli di tempo 157
dati Triple-S
 importazione 26
decimali
 formati di visualizzazione 121
Definizione dei provider di dati 8
definizione dei riquadri 166
definizione della densità in
 Space-Time-Box 93
delimitatori 20, 323
densità
 3-D 176
densità 3D 176
derivazione multipla 128
deviazione standard
 nodo Calcola globali 301
 nodo Raccolta 142
 output del nodo Statistiche 295
deviazione standard/media
 utilizzato per la discretizzazione dei
 campi 142
deviazione standard per
 aggregazione 69
dimensione commit 323
diminuzione di dati 62, 63
direzione dei campi 119
discretizzazione ottimale 142
distribution 215
diversi input 74
documentazione 3
documenti MDD
 importazione 26
DPD (Definizione dei provider di dati) 8
duplicati
 campi 74, 122
 record 80

E

editor query
 nodo origine del database 18, 19
elaborazione parallela
 nodo Aggregazione 69
 ordinamento 73
 unione 79
elementi di tempo ciclico
 preparazione automatica dati 99
elementi grafici
 conversione 263
 modifica 263
elementi grafici (*Continua*)
 modificatori di collisione 263
elementi output 309
eliminazione
 file delle mappe 196
 fogli di stile di visualizzazione 196
 modelli di visualizzazione 196
 oggetti di output 272
entrate
 grafici di valutazione 238
esclusione dei campi inutilizzati
 preparazione automatica dati 99
esecuzione
 definizione dell'ordine di 357
esempi
 cenni generali 5
 Guida alle applicazioni 3
esempi di applicazioni 3
esplorazione
 esplorazione dei dati iniziale 284
 nodo Esplora 284
esplorazione dei dati
 nodo Esplora 284
esplorazione di grafici 243
 aree 248
 bacchetta magica 250
 contrassegno di elementi 250
 sezioni di grafici 244
Esporta numero di decimali 121
esportazione
 dati da IBM Cognos TM1 335
 file delle mappe 196
 fogli di stile di visualizzazione 196
 modelli di visualizzazione 196
 output 274
 Supernodi 358
esportazione di Analytic Server 332
esportazione di dati
 a Excel 337
 file DAT 337
 formato file flat 331
 formato SAS 336
 formato XML 337
 in IBM SPSS Statistics 349
 in un database 316
 Nodo di esportazione IBM Cognos
 BI 37, 333, 334
 nodo di esportazione IBM Cognos
 TM1 335
 testo 337
espressioni CLEM 61
estensione
 campo derivato 128
etichette 118
 esportazione 336, 349
 importazione 39, 342
 specifiche 116, 117, 118
etichette dei valori
 nodo File Statistics 342
etichette di variabile
 nodo Esporta Statistics 349
 nodo File Statistics 342
events
 creazione 231
Excel
 avvio da IBM SPSS Modeler 337

F

- fattori di bilanciamento 68
- fattori di scala 68
- file .sav 342
- file .sd2 (SAS) 39
- file .slb 358
- file .ssd (SAS) 39
- file .tpt (SAS) 39
- file .zsav 342
- file DAT
 - esportazione 274, 337
 - salvataggio 276
- file delimitati da virgole
 - esportazione 274, 337
 - salvataggio 276
- file delle mappe
 - eliminazione 196
 - esportazione 196
 - importazione 196
 - posizione 196
 - ridenominazione 196
 - selezione in Selezionatore modelli per lavagna grafica 174
- file di dati employee_data.sav 343
- file di dati IBM SPSS Statistics
 - importazione dati di indagine 26
- file di Excel
 - esportazione 337
- file di formato 39
- file di output
 - salvataggio 276
- file di testo 20
 - esportazione 337
- file di trasporto
 - nodo origine SAS 39
- File ESRI 197
- file flat 20
- file SMZ
 - cenni generali 197
 - creazione 197
 - eliminazione 196
 - esportazione 196
 - importazione 196
 - modifica di file SMZ preinstallati 197
 - preinstallati 197
 - ridenominazione 196
- file XLS
 - esportazione 337
- filtro di campi 77, 122
 - per IBM SPSS Statistics 350
- finestra del browser
 - Menu Genera 279
- fogli di lavoro
 - importazione da Excel 40
- fogli di stile
 - eliminazione 196
 - esportazione 196
 - importazione 196
 - ridenominazione 196
- fogli di stile di visualizzazione
 - applicazione 267
 - eliminazione 196
 - esportazione 196
 - importazione 196
 - posizione 196
 - ridenominazione 196

- formati
 - dati 24
- formati di archiviazione 24
- formati di ora 121
- formati di output 276
- formati di visualizzazione
 - decimali 121
 - numeri 121
 - scientifico 121
 - simbolo di raggruppamento 121
 - valuta 121
- formati di visualizzazione numeri 121
- formato di visualizzazione
 - scientifico 121
- formato di visualizzazione valute 121
- formato HDATA
 - nodo di origine IBM SPSS Data Collection 25
- formato VDATA
 - nodo di origine IBM SPSS Data Collection 25
- formula di derivazione dei campi 129
- forzatura dei valori 119
- frequenze
 - nodo Raccolta 140
- funzione Max
 - aggregazione serie temporali 153
- funzione Media
 - aggregazione serie temporali 153
- funzione Min
 - aggregazione serie temporali 153
- funzione Moda
 - aggregazione serie temporali 153
- Funzione più recente
 - padding di serie temporali 153
- funzione Primi
 - aggregazione serie temporali 153
- funzione Somma
 - aggregazione serie temporali 153
- funzione Ultimo
 - aggregazione serie temporali 153
- funzione Vero se uno qualsiasi è vero
 - aggregazione serie temporali 153
- funzione Vuoto
 - padding di serie temporali 153
- funzioni di trasferimento 90
 - ordini della differenza 90
 - ordini denominatore 90
 - ordini numeratore 90
 - ordini stagionali 90
 - ritardo 90

G

- Generatore espressioni 61
- generazioni di nodi da grafici 251
 - nodi bilanciamento 251
 - nodi Filtro 251
 - nodi Ricava 251
 - nodi Ricodifica 251
 - nodi Seleziona 251
- gestione dei vuoti 116
 - nodo Raccolta 139
 - riempimento di valori 132
- gestione valori mancanti 95
- gradi di libertà
 - nodo Matrice 279

- gradi di libertà (*Continua*)
 - nodo Medie 298
- grafici
 - 3-D 168
 - aree 248
 - copia 268
 - da Lavagna grafica 170
 - dimensioni degli elementi grafici 257
 - distribuzioni 211
 - eliminazione di aree 248
 - esplorazione 243
 - esportazione 268
 - etichette dell'asse 266
 - foglio di stile 267
 - generazione da un'esplorazione dati 290
 - generazione di nodi 251
 - grafici 205
 - grafici di valutazione 233
 - istogrammi 215
 - multiplot 220
 - nota a piè di pagina 266
 - raccolte 217
 - rotazione di un'immagine 3-D 168
 - salvataggio 268
 - salvataggio dei layout modificati 267
 - salvataggio dell'output 276
 - salvataggio delle modifiche al layout 267
 - scheda Annotazioni 168
 - schede output 167
 - schema di colori di default 267
 - serie temporali 231
 - sezioni 244
 - stampa 268
 - title 266
 - Web 223
- Grafici: opzioni 311
- grafici 3-D 168
- grafici a dispersione 205, 220
- grafici a linee 205, 220
- grafici a punti 205, 220
- grafici dei profitti 233, 241
- Grafici dei profitti 233, 241
- grafici delle risposte 233, 241
- grafici di associazione 223
- grafici guadagno cumulativo 233, 241
- grafico a barre 176
 - 3-D 176
 - dei conteggi 176
 - esempio 184, 185
 - su una mappa 176
- grafico a barre 3D 176
- grafico a barre raggruppato
 - esempio 185
- grafico a bolle 176
- grafico a dispersione 176
 - 3-D 176
 - in bin 176
 - in bin esagonali 176
- grafico a dispersione 3D 176
- grafico a dispersione in bin 176
 - bin esagonali 176
- grafico a dispersione in bin esagonali 176
- grafico a dispersione su matrice (SPLOM) 176

- grafico a nastro 176
- grafico a punti 176
 - 2-D 176
 - esempio 187
- grafico a punti 2D 176
- grafico a scatole 176
 - esempio 188
- grafico a superficie 176
- grafico a torta 176
 - 3-D 176
 - esempio 189
 - su una mappa 176
 - utilizzo dei conteggi 176
- grafico a torta 3D 176
- grafico ad area 176
 - 3-D 176
- grafico ad area 3D
 - descrizione 176
- grafico del percorso 176
- grafico delle coordinate parallele 176
- grafico lineare 176
 - su una mappa 176
- grafico massimo-minimo 176
- grafico massimo-minimo-chiusura 176

H

- hassubstring, funzione 130
- HTML
 - salvataggio dell'output 276

I

- IBM SPSS Collaboration and Deployment Services Repository
 - connessione a 8
 - utilizzo come posizione per modelli, fogli di stile di visualizzazione e mappe 196
- IBM SPSS Modeler 1
 - documentazione 3
- IBM SPSS Modeler Server 1
- IBM SPSS Statistics
 - avvio da IBM SPSS Modeler 312, 346, 349
 - nomi di campi validi 350
 - posizione della licenza 312
- icone, IBM Cognos BI 35
- importanza
 - confronto di medie 297
 - nodo Medie 298
- importazione
 - dati da IBM Cognos BI 36
 - dati da IBM Cognos TM1 38
 - file delle mappe 196
 - fogli di stile di visualizzazione 196
 - modelli di visualizzazione 196
 - report da IBM Cognos BI 36
 - Supernodi 358
- Imposta seme random
 - campionamento di record 146
- impostazione automatica 112, 115
- impostazioni automatiche 255
- incrementi di minuti
 - nodo Intervalli di tempo 160, 161

- incrementi di secondi
 - nodo Intervalli di tempo 161, 162
- indici BITMAP
 - tabelle di database 321
- indicizzazione delle tabelle di database 321
- insiemi
 - conversione in flag 147, 148
 - trasformazione 136, 137
- insiemi a categorie multiple 124
- insiemi a risposta multipla
 - definizione 124
 - eliminazione 124
 - insiemi a categorie multiple 124
 - insiemi di dicotomie multiple 124
 - nelle visualizzazioni 172
 - nodo di origine IBM SPSS Data Collection 25, 26, 29
 - nodo origine IBM SPSS Statistics 342
- insiemi di dicotomie multiple 124
- integrazione
 - modelli ARIMA 89
- intervalli 112
 - dati di serie temporali 151
 - valori mancanti 116
- intervalli di celle
 - file di Excel 40
- intervalli di confidenza
 - nodo Medie 298
- intervalli di numeri interi 117
- intervalli di numeri reali 117
- intervallo
 - output del nodo Statistiche 295
- interventi
 - creazione 231
- istanziamento
 - nodi origine 60
- istogramma 176
 - 3-D 176
 - esempio 186
- istogramma 3D 176
- istruzioni Se, Allora, Altrimenti 131

J

- jitter 208, 254, 263

L

- lavagna grafica
 - tipi di grafici 176
- layout diretto per grafici Web 225
- layout reticolare per grafici Web 225
- legenda
 - posizione 266
- linea base
 - opzioni relative al nodo Valutazione 238
- linea migliore
 - opzioni relative al nodo Valutazione 238
- lingua
 - nodo di origine IBM SPSS Data Collection 28
- lista di campionamento 63

- livellamento esponenziale
 - criteri nei modelli di serie temporali 88
- livellamento LOESS
 - nodo Plot 207
- livellamento LOWESS vedere livellamento LOESS
- livellamento LOESS
 - nodo Plot 207
- livello di misurazione
 - definizione 112
 - modifiche in visualizzazioni 170
 - nelle visualizzazioni 172

M

- manager
 - scheda output 272
- mappa
 - colore 176
 - con frecce 176
 - con grafici a barre 176
 - con grafici a linee 176
 - con grafici a torta 176
 - con punti 176
 - sovrapposizione 176
- mappa colori 176
 - esempio 192
- mappa del flusso 176
- mappa delle coordinate 176
- mappa di sovrapposizione 176
- mappa termica 176
 - esempio 190
- mappatura di campi 317
- mappe
 - assottigliamento 199, 200
 - conversione shapefile ESRI 197
 - distribuzione 204
 - eliminazione caratteristiche 203
 - eliminazione singoli elementi 203
 - etichette caratteristiche 201
 - livellamento 199, 200
 - proiezione 203
 - spostamento caratteristiche 203
 - unione caratteristiche 202
- mapping
 - dati per esportazione a IBM Cognos TM1 336
- mascheramento dei dati da utilizzare in un modello 133
- massimo
 - nodo Calcola globali 301
 - output del nodo Statistiche 295
- matrice di coincidenza
 - nodo Analisi 280
- matrice di grafici a dispersione
 - esempio 191, 193
- media
 - nodo Calcola globali 301
 - nodo Raccolta 142
 - output del nodo Statistiche 295
- Media della funzione più recente
 - padding di serie temporali 153
- media mobile
 - modelli ARIMA 89
- mediana
 - output del nodo Statistiche 295

- medie
 - confronto 296, 297
- membro (importazione SAS)
 - impostazioni 39
- metadati 116
 - nodo di origine IBM SPSS Data Collection 25, 26
- metodo chiave 74
- minimo
 - nodo Calcola globali 301
 - output del nodo Statistiche 295
- misurazioni orarie
 - nodo Intervalli di tempo 159, 160
- misurazioni quotidiane
 - nodo Intervalli di tempo 158, 159
- mode
 - output del nodo Statistiche 295
- modeler Expert
 - criteri nei modelli di serie temporali 87
- modelli
 - anonimizzazione dei dati per 133
 - ARIMA 89
 - eliminazione 196
 - esportazione 196
 - importazione 196
 - nodo Report 299
 - ridenominazione 196
- modelli ARIMA
 - costante 89
 - criteri nei modelli di serie temporali 89
 - funzioni di trasferimento 90
 - ordini autoregressivi 89
 - ordini di differenziazione 89
 - ordini di media mobile 89
 - ordini stagionali 89
 - valori anomali 91
- modelli di serie temporali
 - criteri ARIMA 89
 - criteri di Expert Modeler 87
 - criteri di livellamento esponenziale 88
 - funzioni di trasferimento 90
 - periodicità 90
 - trasformazione della serie 90
 - valori anomali 91
- modelli di visualizzazione
 - eliminazione 196
 - esportazione 196
 - importazione 196
 - posizione 196
 - ridenominazione 196
- modelli IBM SPSS Statistics 346
 - dettagli nugget avanzati 346
 - informazioni su 346
 - nugget del modello 346
 - opzioni modello 346
- modello di elaborazione CRISP-DM Data Preparation 95
- modifica dei grafici
 - dimensioni degli elementi grafici 257
- modifica dei valori dei dati 127
- modifica delle visualizzazioni 254
 - aggiunta di effetti 3-D 262
 - assi 259
 - categorie 261

- modifica delle visualizzazioni (*Continua*)
 - colori e motivi 256
 - combinazione categorie 261
 - compressione delle categorie 261
 - esclusione di categorie 261
 - forma dei punti 257
 - formati dei numeri 258
 - impostazioni automatiche 255
 - margini 258
 - ordinamento di categorie 261
 - padding 258
 - posizione della legenda 266
 - proporzioni dei punti 257
 - regole 255
 - riquadri 262
 - rotazione dei punti 257
 - scale 259
 - selezione 255
 - testo 256
 - trasformazione dei sistemi di coordinate 262
 - trasparenza 256
 - trasposizione 262
 - tratteggio 256
- modificatori di collisione 263

N

- navigazione 307
- nodi di esportazione 315
 - esportazione di Analytic Server 332
- nodi di output 280
- nodi Grafici 165
 - animazione 166
 - Distribuzione 211
 - Grafico 205
 - Histogram 215
 - Lavagna grafica 170
 - Multiplot 220
 - Plot tempo 231
 - Raccolta 217
 - riquadri 166
 - sovrapposizioni 166
 - Valutazione 233
 - Web 223
- nodi IBM SPSS Statistics 341
- Nodi incapsulati 352
- nodi Operazioni su campi 95
 - generazione da un'esplorazione dati 290
- nodi Operazioni su record 61
 - nodo Intervalli di tempo 151
- nodi origine
 - cenni generali 7
 - istanziamento di tipi 60
 - nodo Enterprise View 8
 - nodo File Statistics 342
 - nodo Genera simulazione 42, 44
 - nodo Input utente 55, 56
 - nodo origine del database 13
 - nodo origine Excel 40
 - Nodo origine IBM Cognos BI 34, 37, 38
 - nodo origine IBM Cognos TM1 38
 - nodo origine SAS 39
 - Nodo origine XML 41
 - nodo Testo fisso 22
- nodi origine (*Continua*)
 - nodo Testo variabile 20
 - origine Analytic Server 30
- nodi output 271, 275, 278, 284, 294, 299, 300, 301, 302, 303, 304, 306, 307, 309, 311, 346
 - pubblica su Web 273
 - scheda output 276
- nodo Accodamento
 - cenni generali 80
 - corrispondenza di campi 80
 - impostazione delle opzioni 80
 - impostazione di tag per i campi 78
- nodo Adattamento simulazione 301
 - adattamento delle distribuzioni 302
 - impostazioni di output 303
 - scheda impostazioni 303
- nodo Aggregazione
 - cenni generali 69
 - elaborazione parallela 69
 - impostazione delle opzioni 69
 - performance 69
- nodo Aggregazione RFM
 - cenni generali 72
 - impostazione delle opzioni 72
- nodo Analisi 280
 - scheda Analisi 280
 - scheda output 276
- nodo Analisi RFM
 - cenni generali 144
 - discretizzazione dei valori 145
 - impostazioni 144
- nodo Anonimizza
 - cenni generali 133
 - creazione di valori anonimizzati 135
 - impostazione delle opzioni 134
- nodo bilanciamento
 - cenni generali 68
 - generazione da grafici 251
 - impostazione delle opzioni 68
- nodo Calcola globali 300
 - scheda impostazioni 301
- nodo Campione
 - campioni casuali 63, 64
 - campioni non casuali 63, 64
 - campioni ponderati 66
 - campioni raggruppati 63, 64, 66
 - campioni sistemati 63, 64
 - campioni stratificati 63, 64, 66, 67
 - Dimensioni dei campioni per gli strati 67
 - lista di campionamento 63
- nodo Crea flag 147
- nodo Cronologia 162
 - cenni generali 162
- nodo del grafico temporale 231
 - scheda Aspetto 232
 - scheda Plot 231
 - utilizzo del grafico 233
- nodo dell'insieme
 - campi di output 125
 - combinazione di punteggi 125
- nodo Deriva
 - generazione da un nodo Discretizza 143
 - generazione dai bin 138
- Nodo di esportazione del database 316

- Nodo di esportazione del database
(*Continua*)
 - mappatura dei campi dati di origine
 - alle colonne del database 317
 - nome di tabella 316
 - opzioni di unione 317
 - scheda Esporta 316
 - schema 318
 - sorgente dati 316
 - tabelle di indicizzazione 321
- nodo di esportazione File flat 331
 - scheda Esporta 331
- Nodo di esportazione IBM Cognos BI 37, 333, 334
- nodo di esportazione IBM Cognos TM1 335
- nodo di origine IBM SPSS Data Collection 25, 26, 29
 - file di metadati 26
 - file di registro 26
 - impostazioni delle connessioni al database 29
 - insiemi a risposta multipla 29
 - lingua 28
 - tipi di etichette 28
- nodo distribuzione 211
 - scheda Aspetto 212
 - scheda Plot 211
 - utilizzo del grafico 212
 - utilizzo della tabella 212
- nodo Elimina duplicati
 - cenni generali 80
 - impostazioni di composito 83, 84
 - impostazioni di ottimizzazione 82
 - ordinamento record 80
- nodo Enterprise View 8
- nodo Esplora 284
 - scheda impostazioni 284
 - scheda output 276
- nodo Esporta IBM SPSS Data Collection 331
- nodo Esporta SAS 336
- nodo Esporta Statistics 349
 - scheda Esporta 349
- nodo Esporta XML 337
- nodo Esportazione da Excel 337
- nodo esportazione IBM Cognos TM1
 - esportazione di dati 335
 - mapping dati esportazione 336
- nodo File cache 342
- nodo File Statistics 342
- nodo Filtro
 - cenni generali 122
 - impostazione delle opzioni 122
 - insiemi a risposta multipla 124
- nodo Genera simulazione
 - cenni generali 42
 - impostazione delle opzioni 44
- nodo Importazione da Excel
 - generazione da output 337
- nodo Input utente
 - cenni generali 55
 - impostazione delle opzioni 56
- nodo Intervalli di tempo 152, 153, 154
 - cenni generali 151
- nodo Istogramma 215
 - scheda Aspetto 216
- nodo Istogramma (*Continua*)
 - scheda Plot 215
 - utilizzo del grafico 216
- nodo Lavagna grafica 170
 - scheda Aspetto 194
- nodo Matrice 278
 - browser di output 279
 - evidenziazione 279
 - ordinamento di righe e colonne 279
 - percentuali di colonna 279
 - percentuali di riga 279
 - scheda Aspetto 279
 - scheda impostazioni 278
 - scheda output 276
 - tavole di contingenza 279
- nodo Medie 296
 - browser di output 297, 298
 - campi accoppiati 297
 - gruppi indipendenti 296
 - importanza 297
 - scheda output 276
- nodo Multiplot 220
 - scheda Aspetto 222
 - scheda Plot 220
 - utilizzo del grafico 222
- nodo Ordina
 - cenni generali 73
 - impostazioni di ottimizzazione 73
- nodo origine del database 13
 - editor query 18, 19
 - query SQL 14
 - selezione di tabelle e visualizzazioni 18
- nodo origine Excel 40
- Nodo origine IBM Cognos BI 34, 37, 38
 - icone 35
 - importazione dati 36
 - importazione report 36
- nodo origine IBM Cognos TM1 38
 - importazione dati 38
- nodo origine Microsoft Excel 40
- nodo origine SAS
 - file .sd2 (SAS) 39
 - file .ssd (SAS) 39
 - file .tpt (SAS) 39
 - file di trasporto 39
- Nodo origine XML 41
- nodo Output IBM SPSS Statistics
 - scheda Output 348
- nodo Output Statistics 346
 - scheda Sintassi 347
- nodo Partizione 146
- nodo Plot 205
 - scheda Aspetto 210
 - scheda Opzioni 208
 - scheda Plot 207
 - utilizzo del grafico 210
- Nodo Preparazione automatica dei dati 97
- nodo Raccolta 217
 - anteprima dei bin 143
 - bin a larghezza fissa 139
 - bin con deviazione standard/media 142
 - cenni generali 138
 - conteggio uguale 140
 - impostazione delle opzioni 139
- nodo Raccolta (*Continua*)
 - ottimale 142
 - ranghi 141
 - scheda Aspetto 218
 - scheda Opzioni 217, 218
 - somme uguali 140
 - utilizzo del grafico 219
- nodo Report 299
 - scheda Modello 299
 - scheda output 276
- nodo Ricava
 - cenni generali 127
 - condizionali 131
 - conteggio 131
 - conversione della classe di archiviazione dei campi 132
 - derivazione multipla 128
 - flag 130
 - formula 129
 - generazione da grafici 251
 - generazione dai collegamenti di un grafico Web 227
 - generazione mediante la preparazione automatica dati 110
 - impostazione delle opzioni 128
 - nominale 130
 - ricodifica di valori 132
 - stato 131
- nodo Ricodifica 136, 137
 - cenni generali 136, 138
 - generazione da un nodo distribuzione 212
- nodo Riempimento
 - cenni generali 132
- nodo Riordina campi 163
 - impostazione delle opzioni 163
 - ordinamento automatico 163
 - ordinamento personalizzato 163
- nodo Riorganizza 148, 149
 - con il nodo Aggregazione 148
- nodo Seleziona
 - cenni generali 62
 - generazione da grafici 251
 - generazione dai collegamenti di un grafico Web 227
- nodo Space-Time-Box
 - cenni generali 92
 - definizione della densità 93
- nodo Statistiche 294
 - correlazioni 294
 - etichette di correlazione 294
 - scheda impostazioni 294
 - scheda output 276
 - statistiche 294
- nodo Streaming TS
 - cenni generali 85
 - opzioni dei campi 86
 - opzioni di distribuzione 91
 - opzioni modello 86
- nodo Tabella 275
 - impostazioni di output 275
 - scheda impostazioni 275
 - scheda output 276
- nodo Testo fisso
 - cenni generali 22
 - impostazione delle opzioni 22
 - riconoscimento automatico data 22

- nodo Testo variabile 20
 - impostazione delle opzioni 20
 - riconoscimento automatico data 20
- nodo Tipo
 - cancellare valori 59
 - cenni generali 111
 - copia dei tipi 120
 - dati continui 117
 - dati nominali 118
 - dati ordinali 118
 - gestione dei vuoti 116
 - impostazione del ruolo di modellazione 119
 - impostazione delle opzioni 112, 114
 - tipo di campo flag 118
- nodo Trasformazioni 291
- nodo Trasformazioni Statistics 343
 - impostazione delle opzioni 343
 - scheda Sintassi 343
 - sintassi consentita 344
- nodo Trasponi 150
 - campi numerici 150
 - campi stringa 150
 - nomi di campo 150
- nodo Unione 74
 - cenni generali 74
 - filtro di campi 77
 - impostazione delle opzioni 76, 77
 - impostazione di tag per i campi 78
 - impostazioni di ottimizzazione 79
- nodo Valutazione 233
 - condizione risultato 240
 - espressione punteggio 240
 - lettura dei risultati 241
 - regola di business 240
 - scheda Aspetto 240
 - scheda Opzioni 240
 - scheda Plot 238
 - utilizzo del grafico 242
- nodo Valutazione simulazione 304, 306, 307, 309, 311
 - impostazioni di output 304
 - scheda impostazioni 304
- nodo Vista dati 11
 - impostazione delle opzioni 12
- nodo Web 223
 - definizione di collegamenti 225
 - dispositivo di scorrimento 227
 - dispositivo di scorrimento dei collegamenti 227
 - modifica del layout 227
 - modifica delle soglie 229
 - modifica di punti 227
 - riepilogo Web 231
 - scheda Aspetto 226
 - scheda Opzioni 225
 - scheda Plot 224
 - utilizzo del grafico 227
- nomi di campo 123
 - anonimizzazione 124
 - esportazione di dati 316, 331, 336, 349
- nomi di variabili
 - esportazione di dati 316, 331, 336, 349
- normalizza target continuo 101, 110

- normalizzazione di valori
 - nodi Grafici 220, 231
- null 116
 - nelle tabelle Matrice 278

O

- ODBC
 - caricamento di massa tramite 323, 324
 - connessione per il nodo di esportazione IBM Cognos BI 334
 - nodo origine del database 13
- ODBC, nodo di esportazione. Vedere nodo di esportazione del database 316
- oggetti di output 272
- opzioni
 - IBM SPSS Statistics 312
- opzioni dei ranghi 141
- opzioni di unione, esportazione database 317
- opzioni modello
 - nodo Modello Statistics 346
- Oracle 13
- ordinamento
 - campi 163
 - campi preordinati 73, 82
 - nodo Elimina duplicati 80
 - record 73
- ordinamento di dati 73, 163
- ordine crescente 73
- ordine decrescente 73
- ordine dei dati di input 78
- ordine di esecuzione
 - specifico 357
- ordine naturale
 - modifica 163
- ordini stagionali
 - modelli ARIMA 89
- origine Analytic Server 30
- origine dati
 - connessioni database 15
- output 306, 307, 309
 - esportazione 274
 - generazione di nuovi nodi da HTML 274
 - salvataggio 272
 - stampa 272
- output del nodo Matrice
 - salvataggio come testo 276
- output grafico 309
- output HTML
 - nodo Report 299
 - visualizzazione in un browser 274
- output tabulare
 - riordinamento delle colonne 275
 - selezione di celle 275
- output XML
 - nodo Report 299

P

- palette
 - nascondere 255
 - spostamento 255
 - visualizzazione 255

- parametri
 - impostazione per Supernodi 355
 - in IBM Cognos BI 38
 - proprietà dei nodi 356
 - Supernodi 355, 356
- parametri flusso 18, 19
- parametri Supernodo 355, 356
- pari merito
 - nodo Raccolta 140
- parola chiave FILLFACTOR
 - indicizzazione delle tabelle di database 321
- parola chiave UNIQUE
 - indicizzazione delle tabelle di database 321
- partizionamento dei dati 146
 - grafici di valutazione 238
 - nodo Analisi 280
- performance
 - dati di campionamento 63
 - nodi Deriva 143
 - nodi Discretizza 143
 - nodo Aggregazione 69
 - ordinamento 73
 - unione 79
- periodi
 - nodo Intervalli di tempo 156
- periodi ciclici
 - nodo Intervalli di tempo 156
- periodicità
 - dati di serie temporali 151
 - Modeler di serie temporali 90
- periodo di stima 154
- pesi
 - grafici di valutazione 238
- piani di accesso ai dati 12
- più campi
 - selezione 129
- preparazione automatica dati
 - analisi dei campi 105
 - campi 98
 - collegamenti tra visualizzazioni 104
 - creazione 102
 - dettagli campo 107
 - dettagli dell'azione 108
 - escludi campi 100
 - esclusione dei campi inutilizzati 99
 - generazione di un nodo Ricava 110
 - impostazioni campo 99
 - input, preparazione 101
 - nomina campi 103
 - normalizza target continuo 101, 110
 - obiettivi 97
 - obiettivi, preparazione 101
 - potere predittivo 106
 - prepara date e ore 99
 - preparazione degli input 101
 - preparazione degli obiettivi 101
 - reimposta visualizzazioni 104
 - riepilogo delle azioni 106
 - riepilogo elaborazione campi 104
 - selezione delle funzioni 102
 - tabella campi 106
 - vista del modello 104
- Programma di utilità di conversione
 - mappe 197, 199

- proprietà
 - nodo 356
- proprietà dei nodi 356
- pubblica su Web 273
- punteggi di propensione
 - bilanciamento dei dati 68
- punteggi di propensione regolata
 - bilanciamento dei dati 68
- punti di divisione
 - nodo Raccolta 138
- Python
 - script di caricamento di massa 323, 324

Q

- qualità dei dati
 - Browser del nodo Esplora 288
- query
 - nodo origine del database 13, 14
- query SQL
 - nodo origine del database 13, 14, 18, 19

R

- raggruppamento dei valori 212
- ranghi frazionari 141
- rappresentazione delle associazioni 223
- record
 - conteggi 69
 - etichette 119
 - lunghezza 22
 - trasposizione 150
 - unione 74
- record composti 83
 - impostazioni personalizzate 84
- record di holdout
 - modelli di serie temporali 154
- record incompleti 76
- record univoci 80
- regola di business
 - opzioni relative al nodo
 - Valutazione 240
- regressione di minimi quadrati ponderati
 - localmente
 - nodo Plot 207
- report
 - salvataggio dell'output 276
- report sulla qualità
 - Browser del nodo Esplora 288
- residui
 - nodo Matrice 279
- ricerca
 - browser del nodo Tabella 277
- ricodifica 136, 138
- ricodifica automatica 136
- riconoscimento automatico data 20, 22
- riconoscimento date 20, 22
- ridenominazione
 - campi per l'esportazione 350
 - file delle mappe 196
 - fogli di stile di visualizzazione 196
 - modelli di visualizzazione 196
- ridenominazione degli oggetti di output 272
- riempimento dei dati di serie temporali 153
- righe vuote
 - file di Excel 40
- riordinamento delle colonne
 - browser del nodo Tabella 275, 277
- ristrutturazione dei dati 148
- risultati
 - opzioni relative al nodo
 - Valutazione 240
- ROI
 - grafici 233, 241
- rotazione di grafici 3-D 168
- ruoli
 - specifici di campi 119
- ruoli di modellazione
 - specifici di campi 119

S

- salvataggio
 - oggetti di output 272, 276
 - output 272
- SAS
 - impostazione delle opzioni di importazione 39
- sblocco di Supernodi 353
- scarto
 - campi 122
- scenario 8
- scheda della sintassi
 - nodo Output Statistics 347
- schema
 - Nodo di esportazione del database 318
- script
 - Supernodi 357
- selezione di righe (casi) 62
- selezione di valori 244, 248, 250
- serie temporali 162
- sezioni nei grafici 244
- shapefile 197
- shapefile mappe
 - concetti 198
 - modifica di mappe SMZ preinstallate 197
 - tipi 198
 - utilizzo con Selezionatore modelli per lavagna grafica 197
- significatività
 - intensità delle correlazioni 294
- simbolo decimale 20
 - formati di visualizzazione numeri 121
 - nodo di esportazione File flat 331
- simbolo di raggruppamento
 - formati di visualizzazione numeri 121
- Sintassi XPath 41
- sistemi di coordinate
 - trasformazione 262
- smoother
 - nodo Plot 207
- soglie
 - visualizzazione di soglie di bin 143
- somma
 - nodo Calcola globali 301
- somma (*Continua*)
 - output del nodo Statistiche 295
- sostituzione di valori di campo 132
- sovrapponi 254, 263
- sovrapposizione di colori nei grafici 166
- sovrapposizione di dimensioni nei grafici 166
- sovrapposizione di forme nei grafici 166
- sovrapposizione di riquadri nei grafici 166
- sovrapposizioni per i grafici 166
- sovrascrittura di tabelle di database 316
- SPLOM 176
 - esempio 191, 193
- stampa dell'output 272
- statistica di valutazione delle prestazioni 280
- statistica F
 - nodo Medie 298
- statistiche
 - modifiche in visualizzazioni 263
 - nodo Esplora 284
 - nodo Matrice 278
- statistiche riassuntive
 - nodo Esplora 284
- supernodi
 - blocco 353
 - caricamento 358
 - creazione 352
 - creazione di cache per 357
 - modifica 354
 - nidificazione 353
 - protezione della password 353, 354
 - salvataggio 358
 - sblocco 353
 - script 357
 - Supernodi di elaborazione 351
 - Supernodi origine 351
 - Supernodi terminali 352
 - tipi di 351
 - zoom avanti 354
- Supernodi 351
 - impostazione dei parametri 355
 - uso di commenti con 355

T

- tabelle
 - salvataggio come testo 276
 - salvataggio dell'output 276
 - unione 74
- tag 74, 78
- tavole di contingenza
 - nodo Matrice 278, 279
- test t
 - campioni accoppiati 297
 - campioni indipendenti 296
 - nodo Medie 296, 297, 298
- testo
 - dati 20, 22
 - delimitati 20
- timestamp 112
- tipi di campo
 - nelle visualizzazioni 172
- tipi di dati 22, 95, 112
 - creazione istanza 114

- tipi di etichette
 - nodo di origine IBM SPSS Data Collection 28
- tipi di grafici
 - lavagna grafica 176
- tipi di variabili
 - nelle visualizzazioni 172
- tipo 24
- tipo di utilizzo 24, 112
- tipo flag 112, 118
- tipo insieme 112
- trasformazione a radice quadrata
 - Modeler di serie temporali 90
- trasformazione differenza
 - modelli ARIMA 89
- trasformazione differenza stagionale
 - modelli ARIMA 89
- trasformazione logaritmica
 - Modeler di serie temporali 90
- trasformazione logaritmica naturale
 - Modeler di serie temporali 90
- trasformazioni
 - ricodifica 136, 138
- trasparenza nei grafici 166
- trasposizione di dati 150
- troncamento dei nomi di campo 122, 123

U

- un errore standard della media
 - output del nodo Statistiche 295
- unione di insiemi di dati 80
- unione esterna 74
- unione in base all'ordine 74
- unione interna 74
- unioni 74, 76
 - esterni parziali 77
- unioni parziali 74, 77

V

- valore chiave per aggregazione 69
- valore conteggio per aggregazione 69
- valore di seme random
 - campionamento di record 146
- valore massimo per aggregazione 69
- valore mediano per aggregazione 69
- valore medio per aggregazione 69
- valore medio per i record 69
- valore minimo per aggregazione 69
- valore p
 - importanza 297
- valore quartile per aggregazione 69
- valore seme
 - campionamento e record 146
- valore varianza per aggregazione 69
- valori
 - etichette di campi e valori 116
 - lettura 115
 - specifici 116
- valori anomali
 - modelli ARIMA 91
 - nei modelli di serie temporali 91
- valori anomali additivi
 - Modeler di serie temporali 91

- valori anomali additivi stagionali
 - Modeler di serie temporali 91
- valori anomali di cambiamento di livello
 - Modeler di serie temporali 91
- valori anomali di tendenza locale
 - Modeler di serie temporali 91
- valori anomali innovativi
 - Modeler di serie temporali 91
- valori anomali transienti
 - Modeler di serie temporali 91
- valori attesi
 - nodo Matrice 279
- valori falsi 118
- valori globali 300
- valori mancanti 95, 116, 118
 - in nodi Aggregazione 69
 - nelle tabelle Matrice 278
- valori mancanti definiti dall'utente
 - nelle tabelle Matrice 278
- valori mancanti di sistema
 - nelle tabelle Matrice 278
- valori non definiti 76
- valori null
 - nelle tabelle Matrice 278
- valori preimpostati, connessione a database 15
- valori sommati 69
- valori veri 118
- valori vuoti
 - nelle tabelle Matrice 278
- valutazione dei modelli 280
- valutazione del modello 233
- Variabili codici
 - nodo di origine IBM SPSS Data Collection 26
- Variabili di sistema
 - nodo di origine IBM SPSS Data Collection 26
- variabili SourceFile
 - nodo di origine IBM SPSS Data Collection 26
- varianza
 - output del nodo Statistiche 295
- virgolette
 - importazione di file di testo 20
 - per esportazione in database 316
- vista del modello
 - nella preparazione automatica dati 104
- viste di dati analitici 12
- visualizzazione
 - grafici e diagrammi 165
 - output HTML in un browser 274
- visualizzazione mappe
 - esempio 192
- visualizzazioni
 - assi 259
 - categorie 261
 - colori e motivi 256
 - copia 266
 - forma dei punti 257
 - formati dei numeri 258
 - margini 258
 - modalità di modifica 254
 - modifica 254
 - padding 258
 - posizione della legenda 266

- visualizzazioni (*Continua*)
 - proporzioni dei punti 257
 - riquadri 261, 262
 - rotazione dei punti 257
 - scale 259
 - testo 256
 - trasformazione dei sistemi di coordinate 262
 - trasparenza 256
 - trasposizione 261, 262
 - tratteggi 256
- visualizzazioni mappe
 - creazione 183
- vuoti
 - nelle tabelle Matrice 278

Z

- Zoom indietro da un Supernodo 354



Printed in Ireland