IBM SPSS Modeler 16 Modellazione Nodi

Nota
Prima di utilizzare queste informazioni e il relativo prodotto, leggere le informazioni in “Note” a pagina 297.

Informazioni sul prodotto

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

Indice

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 favorisce l'applicazione di una competenza di business specifica da parte degli utenti, grazie alla quale sarà possibile ottenere modelli di previsione più efficaci ed una riduzione nei tempi di sviluppo delle soluzione. SPSS Modeler offre una vasta gamma di tecniche di creazione di modelli, quali previsione, classificazione, segmentazione ed algoritmi per l'individuazione 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 attività quotidiane, le aziende diventano imprese in grado di effettuare previsioni e di gestire e automatizzare le decisioni, per raggiungere gli obiettivi di business e vantaggi tangibili sulla 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 contattare il supporto tecnico, visitare 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 servizioIBM 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 Modelercome 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. Introduzione alla modellazione

Un modello è un insieme di regole, formule o equazioni che è possibile utilizzare per prevedere un risultato in base a un insieme di campi o di variabili di input. Per esempio, un istituto finanziario potrebbe utilizzare un modello per prevedere la probabilità che i clienti che richiedono un prestito abbiano o meno problemi di insolvenza in base alle informazioni relative a passati clienti già in suo possesso.

La capacità di prevedere un risultato è l'obiettivo principale dell'analisi predittiva e la comprensione del processo di modellazione è essenziale per poter utilizzare IBM SPSS Modeler.

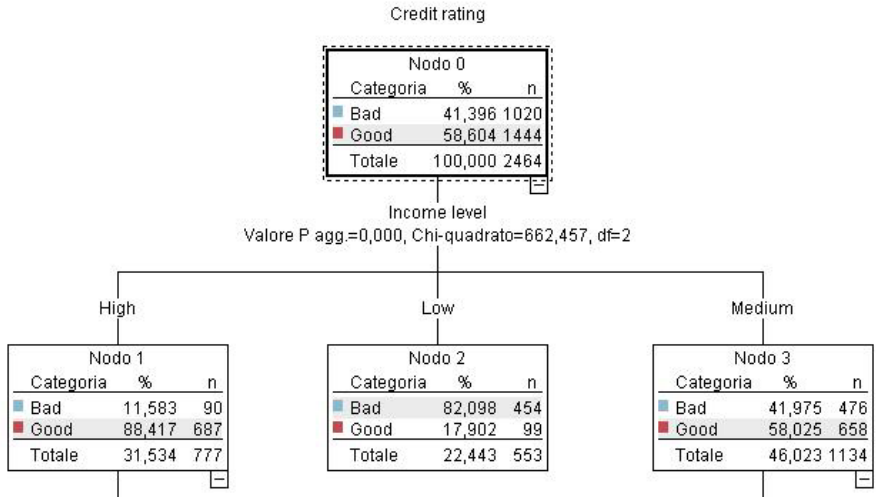


Figura 1. Modello di struttura ad albero delle decisioni semplice

In questo esempio viene utilizzato un modello **Struttura ad albero delle decisioni** che classifica i record (e prevede una risposta) mediante una serie di regole decisionali, per esempio:

```
IF income = Medium
AND cards <5
THEN -> 'Good'
```

Sebbene utilizzi un modello CHAID (Chi-squared Automatic Interaction Detection), questo esempio ha unicamente scopo introduttivo e gran parte dei concetti descritti sono validi in generale anche per gli altri tipi di modellazione in IBM SPSS Modeler.

Per comprendere qualsiasi modello è necessario innanzitutto capire i dati inseriti nel modello. I dati di questo esempio contengono informazioni relative ai clienti di una banca. Vengono utilizzati i campi seguenti:

Nome del campo	**Descrizione**
Credit_rating	Rischio creditizio: 0=Sfavorevole, 1=Favorevole, 9=valori mancanti
Età	Età in anni
Reddito	Livello di reddito: 1=Basso, 2=Medio, 3=Alto
Credit_cards	Numero di carte di credito: 1=Meno di cinque, 2=Cinque o più
Istruzione	Livello di istruzione: 1=Scuola superiore, 2=Università
Car_loans	Numero di mutui auto accesi: 1=Nessuno o uno, 2=Più di due

La banca gestisce un database di informazioni storiche sui clienti che hanno contratto prestiti, inclusi i dati relativi all'avvenuta restituzione (Rischio creditizio = Favorevole) o insolvenza (Rischio creditizio = Sfavorevole). Con i dati esistenti, la banca desidera creare un modello che le consentirà di prevedere la probabilità di insolvenza dei clienti che richiederanno prestiti in futuro.

Mediante un modello di struttura ad albero delle decisioni è possibile analizzare le caratteristiche dei due gruppi di clienti e prevedere la probabilità di mancata restituzione del prestito.

In questo esempio viene utilizzato il flusso denominato *modelingintro.str*, disponibile nella sottocartella *streams* della cartella *Demos*. Il file di dati è *tree_credit.sav*. Per ulteriori informazioni, consultare l'argomento "Cartella Demos" a pagina 5.

Si osservi il flusso.

1. Dal menu principale scegliere:
   File > **Apri flusso**
2. Fare clic sull'icona del nugget dorato nella barra degli strumenti della finestra di dialogo Apri e scegliere la cartella Demos.
3. Fare doppio clic sulla cartella *streams*.
4. Fare doppio clic sul file denominato *modelingintro.str*.

Creazione del flusso

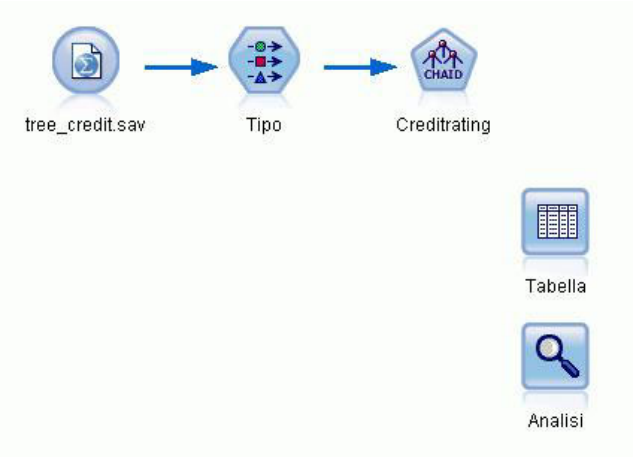


Figura 2. Flusso di modellazione

Per generare un flusso per la creazione di un modello, è necessario disporre di almeno tre elementi:

- Un nodo origine che legge i dati da una sorgente esterna, in questo caso un file di dati IBM SPSS Statistics.
- Un nodo origine o un nodo Tipo che specifica le proprietà dei campi, per esempio il livello di misurazione (il tipo di dati contenuto nel campo) e il ruolo di ogni campo (obiettivo o input) nella modellazione.
- Un nodo Modelli che genera un nugget del modello quando viene eseguito il flusso.

In questo esempio, viene utilizzato un nodo di modellazione CHAID. CHAID, o Chi-squared Automatic Interaction Detection, è un metodo di classificazione che crea strutture ad albero delle decisioni utilizzando un particolare tipo di statistica, noto come statistica chi-quadrato, per calcolare i punti migliori in cui eseguire le suddivisioni nella struttura ad albero delle decisioni.

Se il nodo origine specifica già i livelli di misurazione, è possibile eliminare il nodo Tipo. Dal punto di vista funzionale, il risultato è identico.

Il flusso è dotato inoltre di nodi Tabella e Analisi che saranno utilizzati per visualizzare i risultati del calcolo del punteggio dopo la creazione e l'aggiunta al flusso del nugget del modello.

Il nodo origine File Statistics legge i dati in formato IBM SPSS Statistics dal file di dati *tree_credit.sav*, installato nella cartella *Demos* (per fare riferimento a questa cartella nell'installazione corrente di IBM SPSS Modeler viene utilizzata una variabile speciale denominata *$CLEO_DEMOS*, al fine di garantire che il percorso sia valido indipendentemente dalla cartella o dalla versione dell'installazione corrente).

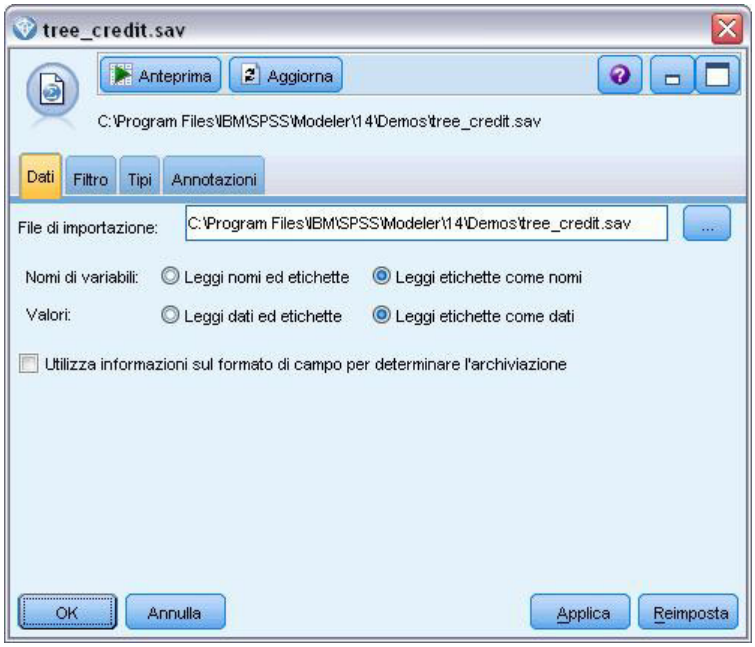


Figura 3. Lettura di dati con un nodo origine File Statistics

Il nodo Tipo specifica il **livello di misurazione** per ogni campo. Il livello di misurazione è una categoria che indica il tipo di dati all'interno del campo. Il file di dati di origine utilizzato in questo esempio impiega tre diversi livelli di misurazione.

Un campo **Continuo** (per esempio il campo *Età*) contiene valori numerici continui, mentre un campo **Nominale** (per esempio il campo *Rischio creditizio*) ha due o più valori distinti, per esempio *Sfavorevole*, *Favorevole* o *Nessuno storico crediti*. Un campo **Ordinale** (come, ad esempio, il campo *Livello di reddito*) descrive i dati con più valori distinti con un ordine intrinseco — in questo caso, *Basso*, *Medio* e *Alto*.

Figura 4. Impostazione dei campi obiettivo e di input con il nodo Tipo

Per ogni campo, il nodo Tipo specifica anche un **ruolo** per indicare il ruolo svolto dai singoli campi nella modellazione. Il ruolo viene impostato su *Obiettivo* per il campo *Rischio creditizio*, ovvero il campo che indica l'insolvenza o meno di un determinato cliente. Questo è l'**obiettivo** o il campo di cui si desidera prevedere il valore.

Per gli altri campi, il ruolo viene impostato su *Input*. I campi di input sono noti anche come **predittori**, o campi i cui valori sono utilizzati dall'algoritmo di modellazione per prevedere il valore del campo obiettivo.

Il nodo Modelli CHAID genera il modello.

Nella scheda Campi del nodo Modelli è selezionata l'opzione **Utilizza ruoli predefiniti** che indica di utilizzare l'obiettivo e gli input specificati nel nodo Tipo. A questo punto è possibile modificare i ruoli dei campi. Nell'esempio, tuttavia, vengono utilizzati i ruoli predefiniti.

1. Fare clic sulla scheda Opzioni di creazione.

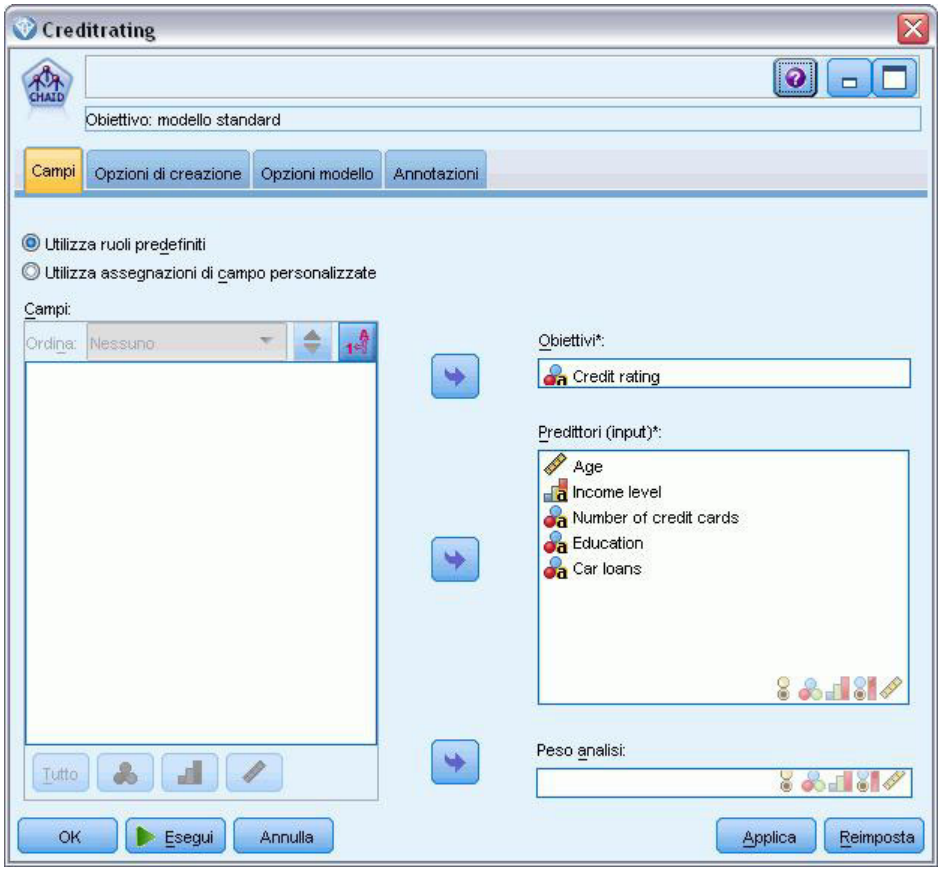


Figura 5. Nodo Modelli CHAID, scheda Campi

La scheda include diverse opzioni che consentono di specificare il tipo di modello che si desidera creare.

Nell'esempio si desidera creare un modello nuovo e si seleziona l'opzione di default **Crea nuovo modello**.

Poiché si desidera creare un solo modello di struttura ad albero delle decisioni standard senza alcuna modifica, si lascia selezionata l'opzione di default **Crea una singola struttura ad albero**.

Anche se è possibile lanciare una sessione di modellazione interattiva che consente di ottimizzare il modello, questo esempio genera semplicemente un modello utilizzando l'impostazione di default **Genera modello**.

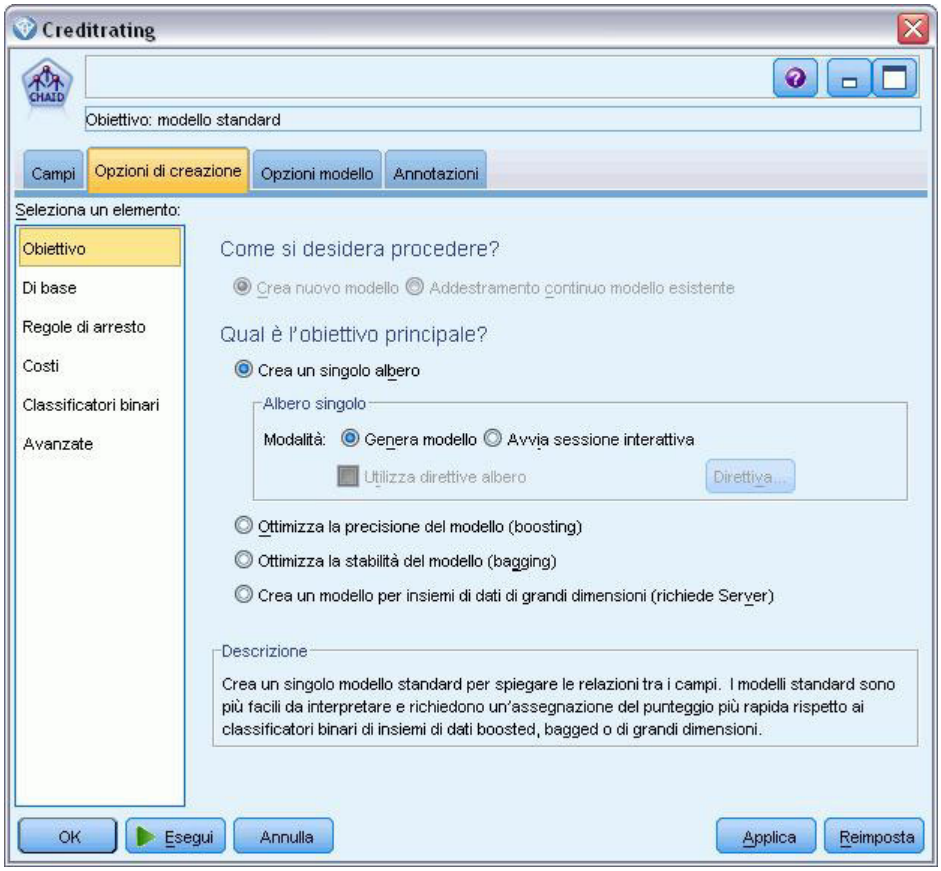


Figura 6. Nodo Modelli CHAID, scheda Opzioni di creazione

Per questo esempio la struttura ad albero dovrà essere abbastanza semplice, per cui se ne limiterà l'espansione aumentando il numero minimo di casi per i nodi padre e figlio.

2. Nella scheda Opzioni di creazione, selezionare **Regole di arresto** dal riquadro di navigazione di sinistra.
3. Selezionare l'opzione **Utilizza valore assoluto**.
4. Impostare **Numero min. record per ramo padre** su 400.
5. Impostare **Numero min. record per ramo figlio** su 200.

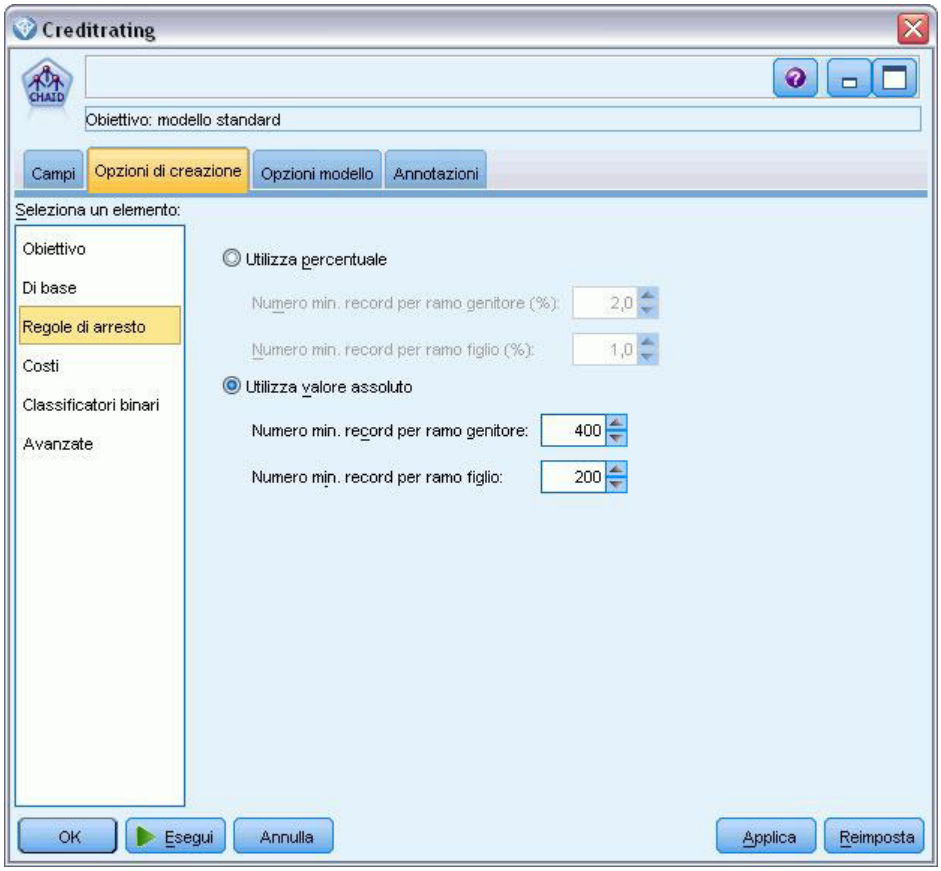


Figura 7. Impostazione dei criteri di arresto per la creazione di strutture ad albero delle decisioni

Poiché è possibile utilizzare tutte le opzioni di default nell'esempio, fare clic su **Esegui** per creare il modello. In alternativa, fare clic sul nodo con il pulsante destro del mouse e scegliere **Esegui** dal menu di scelta rapida oppure selezionare il nodo e scegliere **Esegui** dal menu Strumenti.

Visualizzazione del modello

Una volta terminata l'esecuzione, il nugget del modello viene aggiunto alla palette Modelli nell'angolo superiore destro della finestra dell'applicazione e anche all'area del flusso con un collegamento al nodo Modelli con cui è stato creato. Per visualizzare i dettagli del modello, fare clic con il pulsante destro del mouse sul nugget del modello e scegliere **Visualizza** (dalla palette Modelli) o **Modifica** (dall'area).

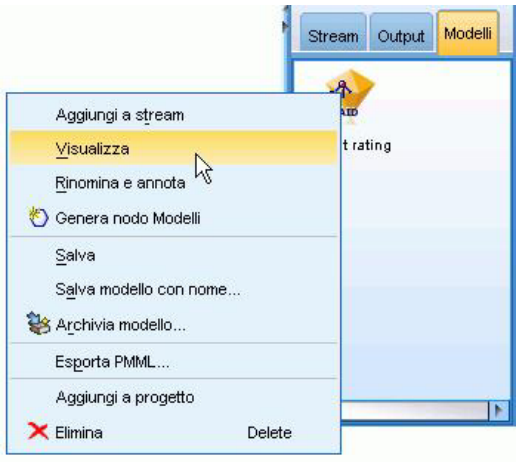


Figura 8. Palette Modelli

Nel caso del nugget CHAID, la scheda Modello visualizza i dettagli sotto forma di insieme di regole ovvero, in sostanza, una serie di regole che è possibile utilizzare per assegnare singoli record a nodi figlio in base ai valori dei vari campi di input.

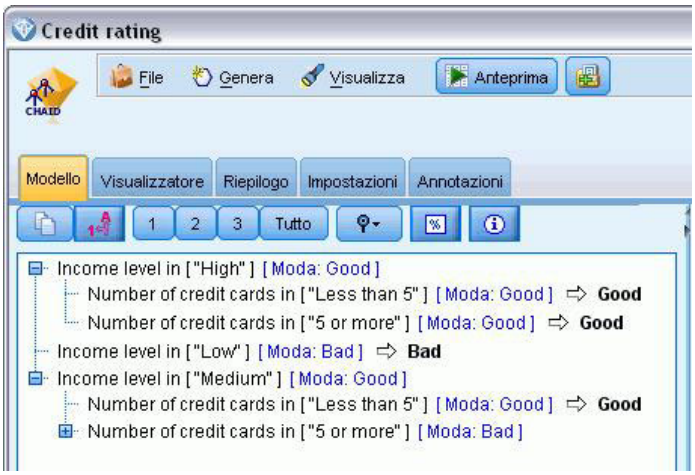


Figura 9. Nugget del modello CHAID, insieme di regole

Per ogni nodo terminale (ovvero nodo che non viene ulteriormente suddiviso) della struttura ad albero delle decisioni viene restituita una previsione *Favorevole* o *Sfavorevole*. Nei singoli casi, la previsione è determinata dalla **moda**, o risposta più comune, dei record che rientrano in quel nodo.

A destra dell'insieme di regole, la scheda Modello visualizza il grafico dell'importanza dei predittori, che visualizza l'importanza relativa di ciascun predittore nella stima del modello. Dal grafico si nota che *Livello di reddito* è probabilmente l'elemento più significativo in questo caso, e che per il resto l'unico fattore significativo è il *Numero di carte di credito*.

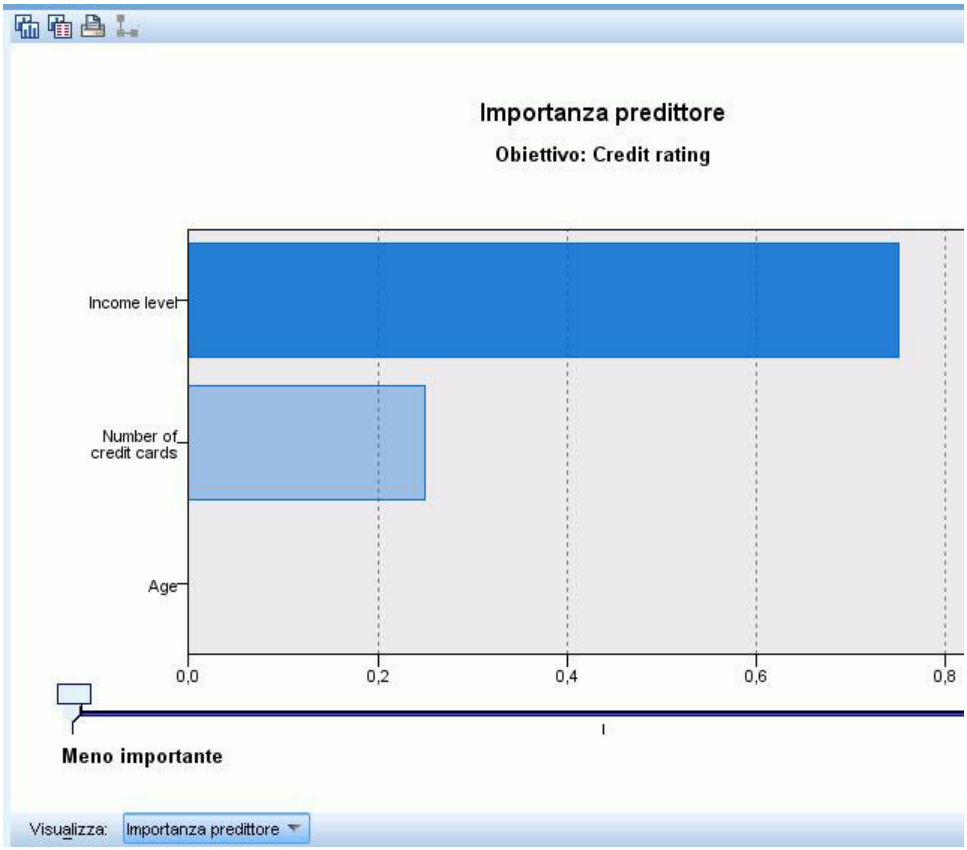


Figura 10. Grafico dell'importanza dei predittori

La scheda Visualizzatore del nugget del modello mostra lo stesso modello sotto forma di struttura ad albero, con un nodo in corrispondenza di ogni punto decisionale. Per ingrandire un determinato nodo o ridurlo in modo da visualizzare meglio la struttura ad albero è possibile utilizzare i comandi di ingrandimento e riduzione.

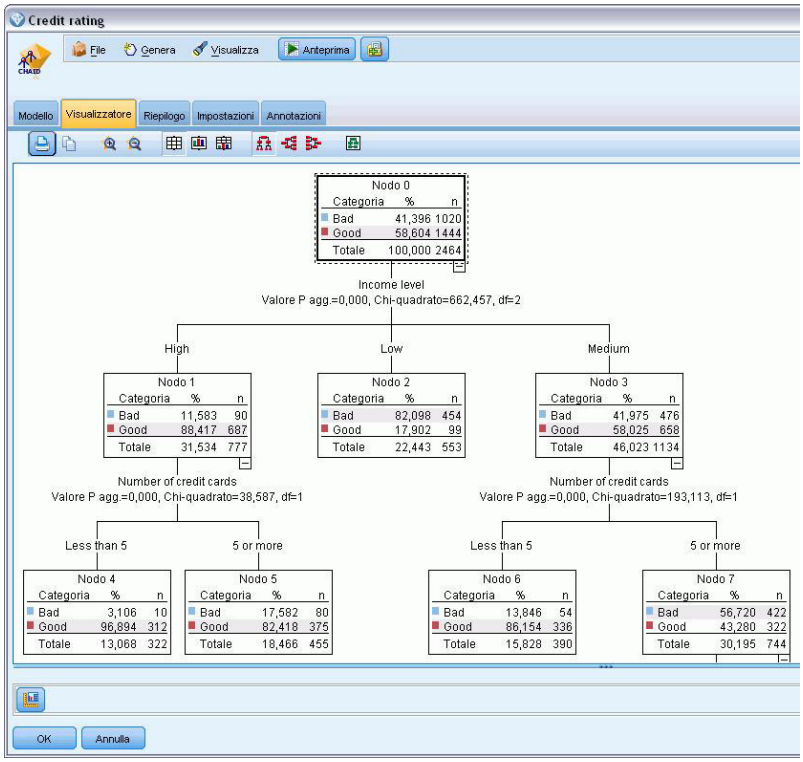


Figura 11. Scheda Visualizzatore del nugget del modello con il comando di riduzione selezionato

Osservando la parte superiore della struttura ad albero, il primo nodo (il Nodo 0) fornisce un riepilogo di tutti i record dell'insieme di dati. Poco più del 40% dei casi dell'insieme di dati è classificato come a rischio creditizio sfavorevole. Questa è una proporzione piuttosto elevata, per cui è necessario esaminare la struttura ad albero per provare ad individuare i fattori responsabili.

Si nota che la prima suddivisione è in corrispondenza di *Livello di reddito*. I record in cui il livello di reddito è presente nella categoria *Basso* sono assegnati al nodo 2 e non sorprende che tale categoria contiene la percentuale più alta di clienti insolventi. Chiaramente, concedere un prestito ai clienti di questa categoria è molto rischioso.

Tuttavia, il 16% dei clienti in questa categoria *non è stato* inadempiente, per cui la previsione non è sempre corretta. Nessun modello è in grado di prevedere tutte le risposte, ma un modello efficace dovrebbe consentire di prevedere la risposta *più probabile* per ciascun record in base ai dati disponibili.

Analogamente, se si analizzano i clienti con reddito alto (Nodo 1), si nota che la maggior parte (l'89%) è a rischio creditizio basso. Tuttavia, anche tra questi clienti più di 1 su 10 è risultato insolvente. È possibile perfezionare i criteri per la concessione dei prestiti in modo da ridurre al minimo il rischio?

Si noti come il modello ha suddiviso questi clienti in due sottocategorie (Nodi 4 e 5) in base al numero di carte di credito possedute. Per i clienti con reddito elevato, se si concede il prestito solo a quelli che sono titolari di meno di 5 carte di credito è possibile aumentare la percentuale di successo dall'89 al 97% - un risultato ancora più soddisfacente.

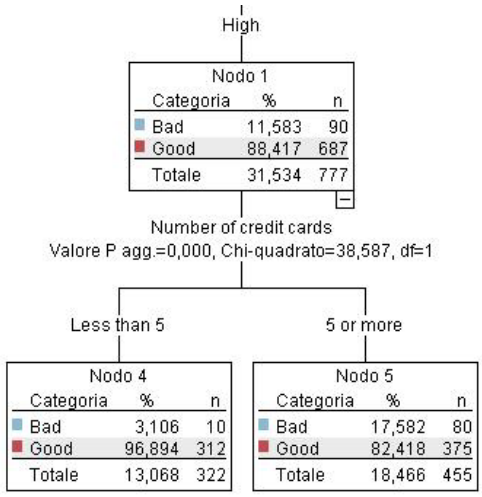


Figura 12. Vista della struttura ad albero dei clienti con reddito elevato

Che cosa accade, però, ai clienti della categoria di reddito Medio (Nodo 3)? Essi sono distribuiti in modo molto più omogeneo tra rischio creditizio favorevole e sfavorevole.

Anche qui, le sottocategorie (Nodi 6 e 7, in questo caso) possono risultare utili. Stavolta, concedere prestiti solo ai clienti con reddito medio titolari di meno di 5 carte di credito aumenta la percentuale di rischio creditizio Favorevole dal 58 all'85%, un miglioramento significativo.

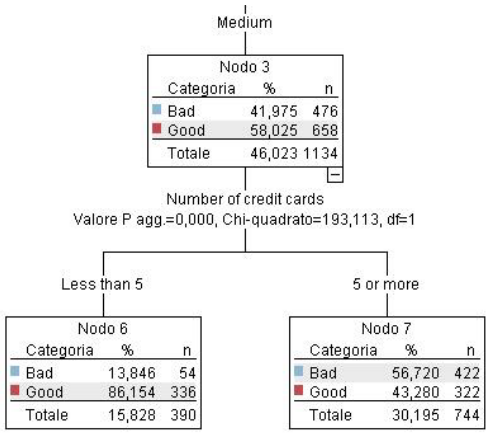


Figura 13. Vista della struttura ad albero dei clienti con reddito medio

Si è visto quindi che ogni record immesso in questo modello viene assegnato ad un nodo specifico e ad esso viene assegnata una previsione *positiva* o *negativa* in base alla risposta più comune per tale nodo.

Questo processo di assegnazione delle previsioni ai singoli record è noto come **calcolo del punteggio**. Calcolando il punteggio degli stessi record utilizzati per la stima del modello, è possibile valutare l'accuratezza dell'esecuzione sui dati di addestramento—i dati per cui si conosce il risultato. Di seguito è illustrato come effettuare tale operazione.

Valutazione del modello

Per comprendere il funzionamento del calcolo del punteggio occorre visualizzare il modello. Per valutare invece il *grado di precisione* del modello è necessario calcolare il punteggio di alcuni record e confrontare le risposte previste dal modello con i risultati effettivi. In questo modo sarà calcolato il punteggio degli stessi record utilizzati per la stima del modello, il che consentirà di confrontare le risposte osservate e

quelle previste.

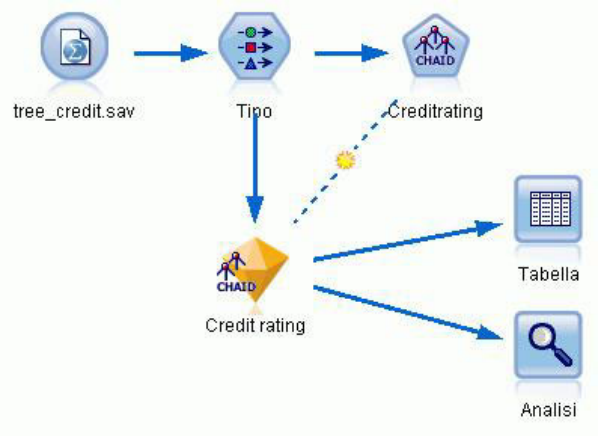
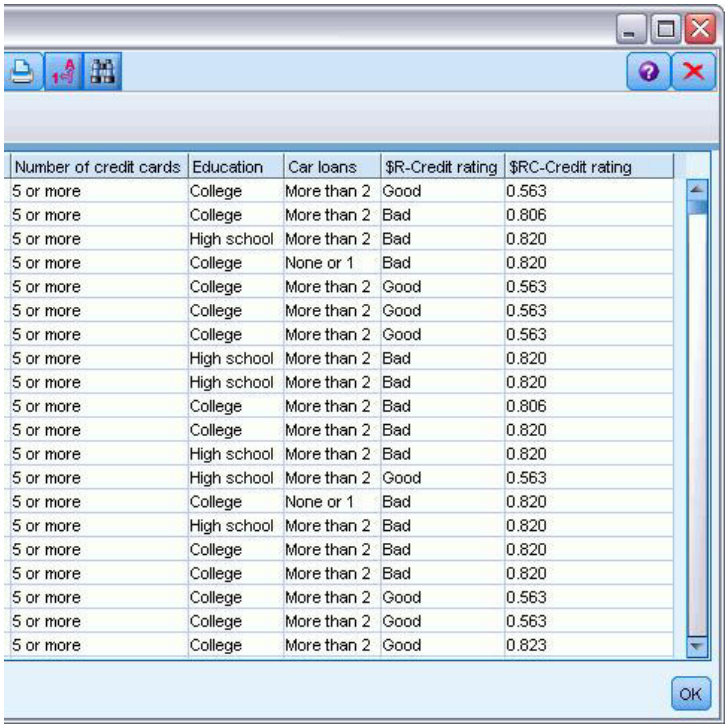


Figura 14. Collegamento del nugget del modello ai nodi di output per la valutazione del modello

1. Per visualizzare i punteggi o le previsioni, associare il nodo Tabella al nugget del modello, fare doppio clic sul nodo Tabella e scegliere **Esegui**.

   La tabella mostra i punteggi previsti in un campo denominato *$R-Rischio creditizio*, generato dal modello. È possibile confrontare tali valori con quelli del campo *Rischio creditizio* originale, contenente le risposte effettive.

   Per convenzione, i nomi dei campi generati durante il calcolo del punteggio hanno lo stesso nome del campo obiettivo, preceduto però da un prefisso standard quale *$R-* per le previsioni o *$RC-* per i valori di confidenza. I vari tipi di modelli adottano insiemi di prefissi diversi. Un **valore di confidenza** è la stima dell'accuratezza di ciascun valore previsto effettuata dal modello, in una scala da 0.0 a 1.0.



Number of credit cards	Education	Car loans	$R-Credit rating	$RC-Credit rating
5 or more	College	More than 2	Good	0.563
5 or more	College	More than 2	Bad	0.806
5 or more	High school	More than 2	Bad	0.820
5 or more	College	None or 1	Bad	0.820
5 or more	College	More than 2	Good	0.563
5 or more	College	More than 2	Good	0.563
5 or more	College	More than 2	Good	0.563
5 or more	High school	More than 2	Bad	0.820
5 or more	High school	More than 2	Bad	0.820
5 or more	College	More than 2	Bad	0.806
5 or more	College	More than 2	Bad	0.820
5 or more	High school	More than 2	Bad	0.820
5 or more	High school	More than 2	Good	0.563
5 or more	College	None or 1	Bad	0.820
5 or more	High school	More than 2	Bad	0.820
5 or more	College	More than 2	Bad	0.820
5 or more	College	More than 2	Bad	0.820
5 or more	College	More than 2	Good	0.563
5 or more	College	More than 2	Good	0.563
5 or more	College	More than 2	Good	0.823

Figura 15. Tabella che mostra i punteggi generati e i valori di confidenza

Come previsto, il valore previsto corrisponde alle risposte effettive per molti record, ma non per tutti, a causa del fatto che ogni nodo terminale CHAID è composto da un insieme eterogeneo di risposte. La previsione corrisponde alla risposta *più comune*, ma risulterà errata per tutte le altre risposte di quel nodo (si rammenti la minoranza del 16% tra i clienti a basso reddito che non è risultata insolvente).

Per evitare questo problema, è possibile continuare a suddividere la struttura ad albero in rami sempre più piccoli, fino a quando ogni nodo non risulta puro al 100% — tutto *Favorevole* o *Sfavorevole*, senza risposte miste. Un modello del genere sarebbe però molto complicato e probabilmente non verrebbe correttamente generalizzato per altri insiemi di dati.

Per determinare con esattezza il numero delle previsioni corrette si potrebbe scorrere la tabella e contare i record in cui il valore del campo previsto *$R-Rischio creditizio* corrisponde al valore di *Rischio creditizio*. Fortunatamente, è disponibile un modo molto più semplice -- è possibile utilizzare un nodo Analisi, che esegue tale operazione automaticamente.

2. Collegare il nugget del modello al nodo Analisi.
3. Fare doppio clic sul nodo Analisi e fare clic su **Esegui**.

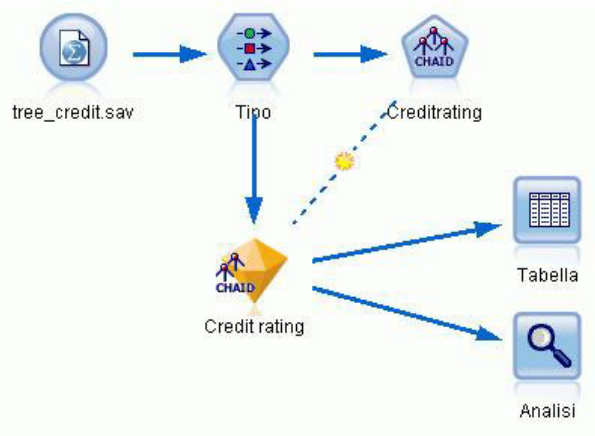


Figura 16. Collegamento di un nodo Analisi

L'analisi mostra che per 1899 record su 2464 (più del 77%) il valore previsto dal modello corrisponde a quello della risposta effettiva.

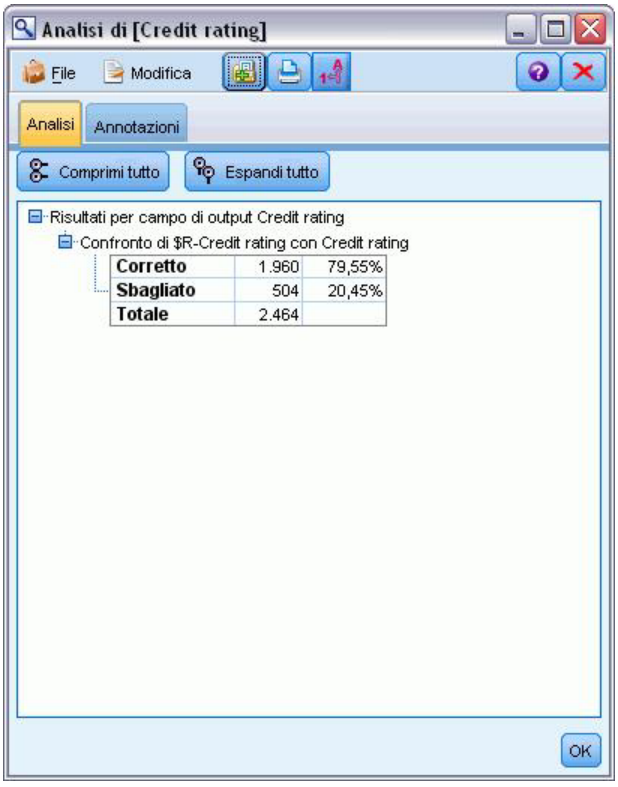


Figura 17. Risultati dell'analisi da confronto tra risposte previste e osservate

Questo risultato è limitato dal fatto che i record di cui si calcola il punteggio sono anche quelli utilizzati per valutare il modello. In una situazione reale si potrebbe utilizzare un nodo Partizione per suddividere i dati in campioni separati per l'addestramento e la valutazione.

Se si utilizza una partizione campione per generare il modello e un altro campione per verificarlo è possibile ottenere un'indicazione molto più precisa della facilità di generalizzazione su altri insiemi di dati.

Il nodo Analisi consente di verificare il modello in relazione ai record per cui il risultato effettivo è già noto. La fase successiva illustra come utilizzare il modello per calcolare il punteggio dei record di cui non si conosce il risultato. Per esempio, il calcolo potrebbe includere le persone che non sono clienti della banca ma che costituiscono i potenziali destinatari di un mailing promozionale.

Calcolo del punteggio dei record

In precedenza è stato calcolato il punteggio degli stessi record utilizzati per la stima del modello al fine di valutare la precisione del modello stesso. Ora si illustrerà come calcolare il punteggio di un insieme di record diverso da quello utilizzato per la creazione del modello. Questo è l'obiettivo della modellazione con un campo obiettivo: studiare i record per cui si conosce il risultato, per identificare i modelli che consentono di prevedere i risultati non ancora noti.

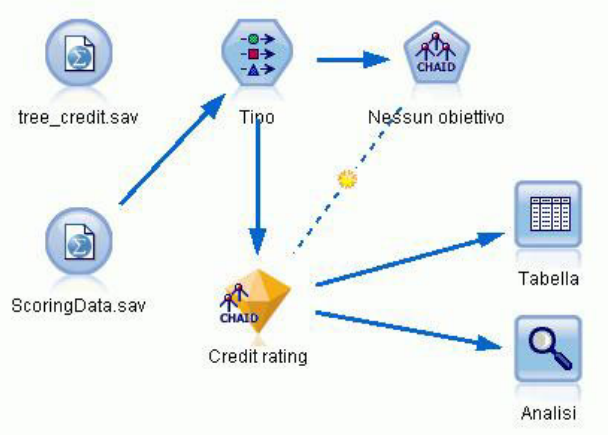


Figura 18. Collegamento di nuovi dati per il calcolo del punteggio

È possibile aggiornare il nodo origine File Statistics in modo che punti a un file di dati diverso, oppure aggiungere un nuovo nodo origine che legga i dati di cui si desidera calcolare il punteggio. Indipendentemente dalla soluzione adottata, il nuovo insieme di dati deve contenere gli stessi campi di input utilizzati dal modello (*Età*, *Livello di reddito*, *Istruzione*, ecc.) ma non il campo obiettivo *Rischio creditizio*.

In alternativa è possibile aggiungere il nugget del modello a qualsiasi flusso che comprenda i campi di input previsti. Che venga letto da un file o da un database, il tipo di sorgente non ha importanza, a condizione che i nomi e i tipi dei campi corrispondano a quelli utilizzati dal modello.

È possibile anche salvare il nugget del modello in un file a parte, esportare il modello in formato PMML per utilizzarlo con altre applicazioni che supportano tale formato o archiviarlo in un repository IBM SPSS Collaboration and Deployment Services, che consente la distribuzione, il calcolo del punteggio e la gestione dei modelli a livello aziendale.

Il modello in sé funziona sempre allo stesso modo, a prescindere dall'infrastruttura utilizzata.

Riepilogo

Questo esempio illustra i passaggi di base per la creazione, la valutazione e il calcolo del punteggio di un modello.

- Il nodo Modelli valuta il modello studiando i record il cui risultato è noto e genera un nugget del modello. Questo processo viene a volte definito "addestramento del modello".
- Il nugget del modello può essere aggiunto a qualsiasi flusso contenente i campi previsti per il calcolo del punteggio dei record. Calcolando il punteggio dei record il cui risultato è già noto (quali quelli dei clienti esistenti) è possibile valutare l'efficacia delle prestazioni dell'insieme di modelli.
- Una volta verificato che le prestazioni del modello sono accettabili, è possibile calcolare il punteggio di nuovi dati (per esempio i potenziali clienti) per prevederne le risposte.
- I dati impiegati per addestrare o valutare il modello vengono talvolta definiti dati analitici o cronologici; i dati per il calcolo del punteggio possono anche venire definiti "dati operativi".

Capitolo 3. Panoramica sulla fase di modellazione

Panoramica sui nodi Modelli

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.

Il documento *IBM SPSS Modeler Guida alle applicazioni* fornisce numerosi esempi di questi metodi, oltre a un'introduzione generale sul processo di modellazione. Tale guida è disponibile come esercitazione online e in formato PDF. Per ulteriori informazioni, consultare “Esempi di applicazioni” a pagina 5.

I metodi di modellazione sono suddivisi in tre categorie:

- Classificazione
- Associazione
- Segmentazione

Modelli Classificazione

I *Modelli Classificazione* utilizzano i valori di uno o più campi di **input** per prevedere il valore di uno o più campi di output o **obiettivo**. Alcuni esempi di queste tecniche sono: strutture ad albero delle decisioni (C&R Tree, QUEST, CHAID e algoritmi C5.0), regressione (lineare, logistica, lineare generalizzata e algoritmo di regressione di Cox), reti neurali, SVM (support vector machines) e reti bayesiane.

I modelli Classificazione consentono di prevedere un risultato noto, per esempio se un cliente effettuerà l'acquisto o vi rinuncerà oppure se una transazione rientra in un modello noto di comportamento fraudolento. Le tecniche di modellazione includono l'apprendimento automatico, l'induzione di regole, l'identificazione dei sottogruppi, i metodi statistici e la creazione di più modelli.

Nodi di classificazione

Il nodo Classificatore automatico crea e confronta svariati tipi di modelli per risultati binari (sì o no, abbandono oppure no e così via), consentendo di scegliere l'approccio migliore per una determinata analisi. Sono supportati numerosi algoritmi di modellazione ed è possibile selezionare i metodi da utilizzare, le opzioni specifiche per ognuno di essi e i criteri per confrontare i risultati. Il nodo genera un insieme di modelli basato sulle opzioni specificate e classifica i candidati migliori in base ai criteri indicati.

Il nodo Numerico automatico stima e confronta i modelli per i risultati di intervalli numerici continui utilizzando svariati metodi. Il nodo funziona in modo analogo al nodo Classificatore automatico e consente di scegliere gli algoritmi da utilizzare e di sperimentare più combinazioni di opzioni in un singolo passaggio di modellazione. Gli algoritmi supportati includono reti neurali, C&R Tree, CHAID, regressione lineare, regressione lineare generalizzata e SVM (Support Vector Machine). I modelli si possono confrontare in base a correlazione, errore relativo o numero di variabili utilizzato.

Il nodo Struttura ad albero di classificazione e regressione (C&R) genera una struttura ad albero delle decisioni che consente di prevedere o classificare osservazioni future. Il metodo utilizza partizionamento ricorsivo per suddividere i record di addestramento in segmenti, riducendo l'impurità ad ogni passaggio. Un nodo della struttura ad albero è considerato "puro" quando il 100% dei casi nel nodo fa parte di una categoria specifica del campo obiettivo. I campi obiettivo e di input possono essere intervalli numerici o categoriali (nominali, ordinali o flag); tutte le suddivisioni sono binarie (solo due sottogruppi).

Il nodo QUEST offre un metodo di classificazione binario per la creazione di strutture ad albero delle decisioni, progettato per ridurre i tempi di elaborazione necessari per le analisi C&R Tree più complesse, riducendo inoltre la tendenza dei metodi per le strutture ad albero di classificazione a favorire gli input che consentono un numero maggiore di suddivisioni. I campi di input possono essere intervalli numerici (continui), ma il campo obiettivo deve essere categoriale. Tutte le suddivisioni sono binarie.

Il nodo CHAID genera una struttura ad albero delle decisioni utilizzando statistiche chi-quadrato per identificare suddivisioni ottimali. A differenza dei nodi C&R Tree e QUEST, il nodo CHAID può generare strutture ad albero non binarie e pertanto alcune suddivisioni possono avere più di due rami. I campi obiettivo e di input possono essere intervallo numerico (continui) o categoriali. Un CHAID completo è una modificazione di CHAID che esegue operazioni avanzate per l'analisi di tutte le suddivisioni possibili, ma richiede tempi di elaborazione maggiori.

Il nodo C5.0 crea una struttura ad albero delle decisioni o un insieme di regole. Il modello suddivide il campione in base al campo che fornisce il massimo guadagno di informazioni a ogni livello. Il campo obiettivo deve essere categoriale. Sono consentite suddivisioni multiple in più di due sottogruppi.

Il nodo Elenco di decisioni identifica i sottogruppi o i segmenti che mostrano una probabilità maggiore o minore che si verifichi un determinato risultato binario rispetto alla popolazione globale. Per esempio, è possibile che si cerchino i clienti non a rischio di abbandono o quelli che più probabilmente rispondano in modo favorevole a una campagna. È possibile incorporare le proprie conoscenze di business nel modello aggiungendo propri segmenti personalizzati e visualizzando in anteprima modelli alternativi uno accanto all'altro per confrontarne i risultati. I modelli Elenco di decisioni consistono in un elenco di regole in cui ogni regola ha una condizione e un risultato. Le regole vengono applicate in ordine e la prima regola corrispondente determina il risultato.

I modelli di regressione lineare prevedono un target continuo basato sulle relazioni lineari tra l'obiettivo e uno o più predittori.

Il nodo fattoriale/PCA offre potenti tecniche di riduzione dei dati che consentono di diminuirne la complessità. L'analisi dei componenti principali (PCA, Principal Components Analysis) trova le combinazioni lineari dei campi di input che catturano meglio la varianza nell'intero insieme di campi, dove i componenti sono ortogonali (perpendicolari) l'uno rispetto all'altro. L'analisi fattoriale tenta di identificare i concetti sottostanti, o fattori, che spiegano lo schema delle correlazioni all'interno dell'insieme di campi osservati. Entrambi gli approcci mirano a trovare un numero ridotto di campi derivati che riassumono in modo efficace le informazioni presenti nell'insieme originale di campi.

Il nodo Selezione funzioni effettua lo screening dei campi di input, rimuovendoli in base a un insieme di criteri quali la percentuale di valori mancanti. Classifica quindi gli input restanti in ordine di importanza rispetto a un determinato obiettivo. Per esempio, dato un insieme di dati con centinaia di input potenziali, quali sono quelli con la maggiore probabilità di essere utili nella modellazione di risultati clinici?

L'analisi discriminante prevede presupposti più rigidi rispetto alla regressione logistica, ma può essere una valida alternativa o un complemento dell'analisi di regressione logistica quando vengono soddisfatti tali presupposti.

La regressione logistica, una tecnica statistica che consente di classificare i record in base ai valori dei campi di input, è analoga alla regressione lineare ma, al posto di un intervallo numerico, prende un campo obiettivo categoriale.

Il modello Lineare generalizzato amplia il modello lineare generale in modo che la variabile dipendente venga linearmente correlata ai fattori e alle covariate tramite una funzione di collegamento specifica. Inoltre, il modello consente alla variabile dipendente di avere una distribuzione non normale. Copre la funzionalità di un grande numero di modelli statistici, inclusi modelli di regressione lineare, modelli di regressione logistica, modelli loglineari per dati dei conteggi e modelli di sopravvivenza censurati per intervallo.

Un modello misto lineare generalizzato (GLMM) estende il modello lineare in modo che l'obiettivo possa avere una distribuzione non normale, sia linearmente correlato ai fattori e alle covariate tramite una funzione di collegamento specifica e in modo che le osservazioni possano essere correlate. I modelli misti lineari generalizzati includono un'ampia gamma di modelli, dalla regressione lineare semplice ai modelli multilivello complessi per i dati longitudinali non normali.

Il nodo Regressione di Cox consente di generare un modello di sopravvivenza per i dati della relazione tempo-evento in presenza di record censurati. Il modello produce una funzione di sopravvivenza che prevede la probabilità che l'evento di interesse si sia verificato a una determinata ora (*t*) per i valori dati delle variabili di input.

Il nodo SVM (Support Vector Machine) consente di classificare i dati in uno di due gruppi senza sovradattamento. Il nodo SVM è particolarmente indicato per l'utilizzo con insiemi di dati di grandi dimensioni, cioè quelli con un elevato numero di campi di input.

Il nodo Rete bayesiana consente di generare un modello di probabilità combinando elementi osservati e registrati con conoscenze del mondo reale per stabilire la probabilità di occorrenze. Il nodo si concentra sulle reti TAN (Tree Augmented Naïve Bayes) e coperta di Markov, che sono prevalentemente utilizzate a scopo di classificazione.

Il nodo Modello risposta autoapprendimento consente di creare un modello in cui è possibile utilizzare un unico nuovo caso oppure un numero limitato di nuovi casi per eseguire una nuova stima del modello senza doverlo riaddestrare con tutti i dati.

Il nodo Serie temporali stima i modelli di livellamento esponenziale, i modelli ARIMA (Autoregressive Integrated Moving Average, autoregressivi integrati a media mobile) univariati e ARIMA (o a funzione di trasferimento) multivariati per i dati di serie temporali e genera previsioni di prestazioni future. Il nodo Serie temporali deve sempre essere preceduto da un nodo Intervalli di tempo.

Il nodo Elemento vicino più prossimo *K* (KNN) associa un nuovo caso alla categoria o valore degli oggetti *K* più vicini ad esso nello spazio predittore, dove *K* è un numero intero. I casi simili sono vicini gli uni agli altri, mentre i casi dissimili sono distanti gli uni dagli altri.

Modelli Associazione

I *Modelli Associazione* trovano nei dati gli schemi nei quali una o più entità (eventi, acquisti o attributi) vengono associate a un'altra o a diverse altre entità. I modelli creano insiemi di regoli che definiscono queste relazioni. I campi all'interno dei dati possono essere sia di input che obiettivo. Queste associazioni possono essere trovate manualmente, ma gli algoritmi delle regole di associazione funzionano più velocemente e possono esplorare schemi più complessi. I modelli Apriori e Carma sono esempi dell'utilizzo di questi algoritmi. Un altro tipo di modello di associazione è un modello di rilevamento di sequenze, che trova gli schemi sequenziali nei dati a struttura temporale.

I modelli Associazione sono utili soprattutto nella previsione di risultati multipli; per esempio, i clienti che hanno acquistato il prodotto X hanno acquistato anche Y e Z. I modelli Associazione associano una particolare conclusione (per esempio la decisione di acquistare qualcosa) a un insieme di condizioni. Rispetto agli algoritmi della struttura ad albero delle decisioni più tradizionali (C5.0 e C&RT), quelli delle regole di associazione presentano il vantaggio di poter definire associazioni tra qualsiasi tipo di attributo. Mentre un algoritmo della struttura ad albero delle decisioni genera regole con un'unica conclusione, gli algoritmi di associazione tentano di individuare più regole, ciascuna delle quali può fornire una diversa conclusione.

Nodi di associazione

Il nodo Apriori estrae un insieme di regole dai dati, estrapolando le regole con il più alto contenuto di informazioni. Apriori offre cinque diversi metodi per la selezione delle regole e utilizza uno schema di indicizzazione sofisticato per elaborare in modo efficiente insiemi di dati di grandi dimensioni. In caso di problemi complessi, l'addestramento di Apriori è in genere più rapido. Apriori non ha un limite arbitrario per quanto riguarda il numero di regole che possono essere mantenute e può gestire regole con un massimo di 32 precondizioni. Apriori richiede che tutti i campi di input e output siano categoriali ma garantisce prestazioni migliori perché è ottimizzato per questo tipo di dati.

Il modello CARMA estrae un insieme di regole dai dati senza che venga richiesto all'utente di specificare i campi di input o obiettivo. A differenza di Apriori , il nodo CARMA fornisce le impostazioni di creazione per il supporto delle regole (sia per l'antecedente che per il conseguente) anziché solo per il supporto antecedente. Pertanto, le regole generate possono essere utilizzate per una gamma più vasta di applicazioni, ad esempio per trovare un elenco di prodotti o di servizi (antecedenti) il cui conseguente è rappresentato dall'articolo che si desidera promuovere per le festività correnti.

Il nodo Sequenza consente di scoprire le regole di associazione nei dati sequenziali o basati su valori temporali. Per sequenza si intende un elenco di serie di elementi che tendono a ricorrere secondo un ordine prevedibile. Ad esempio, un cliente che acquista un rasoio e la lozione dopobarba potrebbe in seguito acquistare la schiuma da barba. Il nodo Sequenza si basa sull'algoritmo delle regole di associazione CARMA, che utilizza un metodo efficiente in due passaggi per trovare le sequenze.

Modelli di segmentazione

I *Modelli di segmentazione* suddividono i dati in segmenti, o cluster, di record con schemi simili di campi di input. I modelli di segmentazione si concentrano soltanto sui campi di input e quindi non tengono in considerazione i campi di output o obiettivo. Tra gli esempi di modelli di segmentazione citiamo le reti di Kohonen, il raggruppamento in cluster Medie K, il raggruppamento in cluster TwoStep e il rilevamento delle anomalie.

I modelli di segmentazione (denominati anche "modelli di cluster") sono utili nei casi in cui il risultato specifico non è conosciuto, per esempio nell'identificazione di nuovi modelli di comportamento fraudolento oppure nell'identificazione di gruppi di interesse nella base clienti. I modelli di raggruppamento tramite cluster sono incentrati sull'identificazione di gruppi di record simili e sull'assegnazione di etichette ai record in base al gruppo di appartenenza. Ciò avviene senza che siano necessarie previe conoscenze relative ai gruppi e alle loro caratteristiche e distingue i modelli di raggruppamento tramite cluster da altre tecniche di modellazione - non esiste un campo di destinazione o un output predefinito per il modello da prevedere. Per questi modelli non esistono risposte corrette o errate. Il valore è determinato dalla capacità di acquisire gruppi significativi all'interno dei dati e di fornire descrizioni utili di tali raggruppamenti. Spesso i modelli di raggruppamento tramite cluster vengono utilizzati per creare cluster o segmenti poi utilizzati come input nelle analisi successive, per esempio segmentando i potenziali clienti in sottogruppi omogenei.

Nodi di segmentazione

Il nodo Cluster automatico stima e confronta i modelli di cluster che identificano gruppi di record con caratteristiche simili. Il nodo funziona in modo analogo ad altri nodi Modelli automatici e consente di sperimentare varie combinazioni di opzioni in un singolo passaggio di modellazione. I modelli si possono confrontare utilizzando misure di base con cui tentare di filtrare e classificare l'utilità dei modelli di cluster e fornire una misura in base all'importanza di determinati campi.

Il nodo Medie K raggruppa l'insieme di dati in gruppi distinti (o cluster). Il metodo definisce un numero fisso di cluster, esegue un'assegnazione iterativa dei record ai cluster e modifica i centri di cluster finché un'ulteriore ridefinizione non consente più un miglioramento del modello. Invece di tentare di prevedere un risultato, il nodo *K*-medie utilizza un processo denominato apprendimento non supervisionato per scoprire gli schemi nell'insieme di campi di input.

Il nodo Kohonen genera un tipo di rete neurale che può essere utilizzato per raggruppare l'insieme di dati in gruppi distinti. Al termine dell'apprendimento della rete, i record analoghi dovranno essere vicini nella mappa di output, mentre i record diversi saranno a notevole distanza. Per identificare le unità forti, è possibile controllare il numero di osservazioni catturate da ciascuna unità nel nugget del modello. In questo modo è possibile avere un'idea del numero appropriato di cluster.

Il nodo TwoStep è un metodo di raggruppamento tramite cluster in due fasi. La prima fase esegue un singolo passaggio nei dati per comprimere i dati di input non elaborati in un insieme gestibile di cluster secondari. Nella seconda fase viene utilizzato un metodo di raggruppamento tramite cluster gerarchico per unire progressivamente i cluster secondari in cluster sempre più grandi. Il nodo TwoStep offre il vantaggio di stimare automaticamente il numero ottimale di cluster per i dati di addestramento. Può gestire in modo efficiente tipi di campo misti e insiemi di dati di grandi dimensioni.

Il nodo Rilevamento anomalie identifica casi insoliti, o valori anomali, non conformi a schemi di dati “normali”. Con questo nodo è possibile identificare valori anomali anche se questi non rientrano in schemi precedentemente conosciuti e anche se l'utente non sa esattamente ciò che sta cercando.

Modelli mining nel database

IBM SPSS Modeler supporta l'integrazione con gli strumenti di data mining e di modellazione offerti dai fornitori di database, quali Oracle Data Miner, IBM DB2 InfoSphere Warehouse e Microsoft Analysis Services. Operando all'interno dell'applicazione IBM SPSS Modeler è infatti possibile sia creare modelli che calcolarne il punteggio e archiviarli nel database. Per ulteriori informazioni, consultare la *IBM SPSS Modeler Guida al mining nel database* disponibile nel DVD del prodotto.

Modelli IBM SPSS Statistics

Se si dispone di una copia di IBM SPSS Statistics installata con licenza sul proprio computer, è possibile accedere a determinate routine IBM SPSS Statistics ed eseguirle dall'interno di IBM SPSS Modeler per creare e determinare il punteggio dei modelli.

Ulteriori informazioni

È inoltre disponibile una documentazione esaustiva sugli algoritmi di modellazione. Per ulteriori informazioni, vedere la *IBM SPSS Modeler Algorithms Guide* disponibile sul DVD del prodotto.

Creazione di modelli di suddivisione

La modellazione suddivisa consente di utilizzare un unico flusso per creare modelli separati per ciascun valore possibile di un campo di input flag, nominale o continuo e i modelli che ne risultano sono tutti accessibili da un unico nugget del modello. I valori possibili per i campi di input potrebbero avere effetti molto diversi sul modello. Con la modellazione suddivisa, è possibile creare facilmente il modello più adatto a ciascun valore di campo possibile in un'unica esecuzione del flusso.

Tenere presente che le sessioni di modellazione interattive non possono utilizzare la suddivisione. Con la modellazione interattiva, ogni modello viene specificato singolarmente, pertanto non c'è alcun vantaggio nell'usare la suddivisione che, invece, crea più modelli automaticamente.

La modellazione suddivisa funziona designando un particolare campo di input come campo di suddivisione. È possibile effettuare tale operazione impostando il ruolo del campo su **Suddividi** nella specifica Tipo.

È possibile designare come campi di suddivisione solo i campi con livello di misurazione **Flag**, **Nominale**, **Ordinale** o **Continuo**.

È possibile assegnare più di un campo di input come campo di suddivisione. In questo caso, tuttavia, il numero di modelli creati potrebbe essere significativamente maggiore. Viene creato un modello per ciascuna possibile combinazione di valori dei campi di suddivisione selezionati. Per esempio, se vengono designati come campi di suddivisione tre campi di input, ognuno dei quali presenta tre valori possibili, questo porterà alla creazione di 27 modelli diversi.

Anche dopo aver assegnato uno o più campi come campi di suddivisione, è possibile scegliere se creare modelli di suddivisione o un modello singolo, mediante l'impostazione di una casella di controllo nella finestra di dialogo del nodo di modellazione.

Se vengono definiti campi di suddivisione, ma non viene selezionata la casella di controllo, verrà generato un unico modello. Analogamente, se la casella di controllo è selezionata, ma non viene definito alcun campo di suddivisione, la suddivisione viene ignorata e viene creato un unico modello.

Quando si esegue il flusso, per ogni possibile valore del campo o dei campi suddivisi vengono generati modelli separati in background, ma nella palette Modelli e nell'area del flusso viene visualizzato un solo nugget del modello. Un nugget del modello di suddivisione è indicato dal simbolo di suddivisione; tale simbolo è composto da due rettangoli grigi sovrapposti all'immagine del nugget.

Quando si visualizza il nugget del modello di suddivisione, viene visualizzato un elenco di tutti i modelli separati che sono stati creati.

È possibile analizzare un singolo modello da un elenco facendo doppio clic sulla relativa icona del nugget nel visualizzatore. In questo modo viene aperta una finestra del browser standard per il singolo modello. Quando il nugget si trova nell'area, il doppio clic sulla miniatura di un grafico apre il grafico nelle dimensioni normali. Per ulteriori informazioni, consultare l'argomento "Visualizzatore del modello di suddivisione" a pagina 47.

Se un modello viene creato come modello di suddivisione, non è possibile rimuovere l'elaborazione della suddivisione, né annullare la suddivisione ancora più a valle da un nodo o un nugget di modellazione suddivisa.

Esempio. Un rivenditore che opera sul territorio nazionale desidera stimare le vendite in base alla categoria di prodotto in ciascuno dei suoi punti vendita sparsi per il paese. Utilizzando la modellazione suddivisa, designa il campo Punto vendita dei dati di input come campo di suddivisione, consentendo di creare modelli separati per ciascuna categoria presso ciascun punto vendita in un'unica operazione. Le informazioni risultanti possono quindi essere utilizzate per controllare i livelli delle scorte in modo molto più preciso rispetto a quanto avviene con un unico modello.

Suddivisione e partizionamento

La suddivisione condivide alcune funzioni con il partizionamento, ma le due operazioni sono utilizzate con modalità diverse.

Il **partizionamento** divide l'insieme di dati in modo casuale in due o tre parti: addestramento, verifica e (come opzione) convalida e viene utilizzato per verificare le prestazioni di un modello singolo.

La **suddivisione** divide l'insieme di dati in tante parti quanti sono i valori possibili per un campo di suddivisione ed è utilizzata per creare più modelli.

Il partizionamento e la suddivisione operano in modo del tutto indipendente. In un nodo Modelli è possibile scegliere una delle due, entrambe o nessuna delle due.

Nodi Modelli che supportano modelli di suddivisione

Alcuni nodi Modelli possono creare modelli di suddivisione. Le eccezioni sono Cluster automatico, Serie temporali, Fattoriale/PCA, Selezione funzioni, SLRM, modelli associazione (Apriori, Carma e Sequenza), i modelli di raggruppamento (Medie K, Kohonen, Two Step ed Anomalia) modello Statistics ed i nodi utilizzati per la modellazione nel database.

I nodi Modelli che supportano la modellazione suddivisa sono:

C&R Tree

Rete di Bayes

QUEST

GenLin

	CHAID		KNN
	C5.0		Cox
	Rete neurale		Classificatore automatico
	Elenco di decisioni		Numerico automatico
	Regressione		Logistico
	Discriminante		SVM

Funzioni su cui ha conseguenze la suddivisione

L'utilizzo dei modelli di suddivisione influisce in vari modi su una serie di funzioni IBM SPSS Modeler. Questa sezione fornisce indicazioni sulle modalità di impiego dei modelli di suddivisione quando utilizzati insieme ad altri nodi in un flusso.

Nodi Oper su record

Quando si utilizzano modelli di suddivisione in un flusso che contiene un nodo **Campione**, stratificare i record per il campo suddiviso al fine di ottenere un campionamento uniforme dei record. Questa opzione è disponibile quando si sceglie il metodo di campionamento complesso.

Se il flusso contiene un nodo **bilanciamento**, tenere presente che il bilanciamento si applica all'insieme complessivo dei record di input e non al sottoinsieme di record in una suddivisione.

Quando si aggregano record per mezzo di un nodo **Aggregazione**, impostare i campi suddivisi in modo che siano campi chiave se si desidera calcolare gli aggregati per ciascuna suddivisione.

Nodi Oper su campi

Il nodo **Tipo** è il punto in cui si specifica il campo o i campi da utilizzare come campi suddivisi.

Tenere presente che, se da un lato il nodo **Insieme** viene utilizzato per combinare due o più nugget del modello, dall'altro non può essere utilizzato per invertire l'azione di suddivisione, in quanto i modelli di suddivisione sono contenuti in un unico nugget del modello.

Nodi modelli

I modelli di suddivisione non supportano il calcolo dell'importanza dei predittori (l'importanza relativa dei campi di input nella stima del modello). Le impostazioni di importanza dei predittori vengono ignorate durante la creazione dei modelli di suddivisione.

Il nodo **KNN** (elemento adiacente più vicino) supporta modelli di suddivisione solo se è impostato per prevedere un campo obiettivo. L'impostazione alternativa (identifica solo gli elementi adiacenti più vicini) non crea un modello. Se si seleziona l'opzione "Seleziona automaticamente K", ogni modello di suddivisione può avere un numero diverso di elementi adiacenti più vicini. Il modello complessivo avrà quindi un numero di colonne generate pari al numero maggiore di elementi adiacenti più vicini trovato in tutti i modelli di suddivisione. Per i modelli di suddivisione in cui il numero di elementi adiacenti più vicini è inferiore a questo valore massimo vi sarà un numero corrispondente di colonne compilate con valori `$null`. Per ulteriori informazioni, consultare l'argomento “Nodo KNN” a pagina 287.

Nodi Modelli database

I nodi di modellazione nel database non supportano i modelli di suddivisione.

Nugget del modello

Non è possibile **esportare in PMML** da un nugget del modello di suddivisione, poiché il nugget contiene modelli multipli e PMML non supporta assemblaggi di questo genere. È tuttavia possibile esportare come testo o HTML.

Opzioni dei campi dei nodi Modelli

In tutti i nodi Modelli è disponibile una scheda Campi nella quale è possibile specificare i campi da utilizzare per la creazione del modello.

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. Le eccezioni riguardano il nodo Sequenza e il nodo Estrazione testo, per i quali è necessario che le impostazioni dei campi vengano specificate nel nodo Modelli.

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.

Nota: non tutti i campi vengono visualizzati per tutti i nodi.

- **Utilizza formato transazionale (solo nodi Apriori, CARMA, Regole di Associazione MS e Oracle Apriori).** Selezionare questa casella di controllo se i dati di origine sono in **formato transazionale**. I record in questo formato hanno due campi, uno per l'ID e uno per il contenuto. Ogni record rappresenta una singola transazione o elemento, e gli elementi associati sono collegati poiché hanno lo stesso ID. Deselezionare questa casella se i dati sono in **formato tabulare**, in cui gli elementi sono rappresentati da flag separati e ogni campo flag rappresenta la presenza o l'assenza di un elemento specifico, mentre ogni record rappresenta un insieme completo di elementi associati. Per ulteriori informazioni, consultare l'argomento “Dati in formato tabellare e dati transazionali” a pagina 230.
  - **ID.** Per i dati transazionali, selezionare un campo ID dall'elenco. Come campo ID è possibile utilizzare campi numerici o simbolici. Ogni valore univoco di questo campo deve indicare una specifica unità di analisi. In un'applicazione per analisi market basket, per esempio, ciascun ID

potrebbe rappresentare un singolo cliente. Per un'applicazione per analisi di registri Web, ciascun ID potrebbe rappresentare un computer (per indirizzo IP) o un utente (per dati di login).

– **ID contigui.** (Solo nodi Apriori e CARMA) Se i dati sono preordinati in modo che tutti i record con lo stesso ID vengano raggruppati 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.

  Nota: se i dati non sono ordinati e viene selezionata questa opzione, potrebbero essere restituiti risultati non validi nel modello.

– **Contenuto.** Specificare il campo o i campi contenuto per il modello. Questi campi contengono gli elementi rilevanti nella creazione di modelli di associazione. È possibile specificare più campi flag se i dati sono in formato tabulare o un singolo campo nominale se i dati sono in formato transazionale.

- **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.
- **Valutazione.** (Solo per modelli Cluster automatico.) Per i modelli Cluster non viene specificato alcun obiettivo; tuttavia, è possibile selezionare un campo di valutazione per identificare il relativo livello di importanza. Inoltre, è possibile valutare la bontà della differenziazione dei valori di questo campo da parte dei cluster, il che, a sua volta, indica se i cluster possono essere utilizzati per prevedere questo campo.
  – **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.
  – **Partizione.** Questo campo consente di specificare un campo utilizzato per partizionare i dati in campioni distinti per le fasi di addestramento, di test e di convalida della creazione del modello. Utilizzando un campione per generare il modello e un altro campione per sottoporlo a verifica, è possibile ottenere un'indicazione valida del modo in cui il modello potrà essere esteso a insiemi di dati di dimensioni maggiori, simili ai dati correnti. Se sono stati definiti più campi partizione utilizzando nodi Tipo o Partizione, 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. Si noti inoltre che per applicare la partizione selezionata nell'analisi, è necessario che il partizionamento sia attivato anche nella scheda Opzioni modello relativa al nodo. La disattivazione di questa opzione consente di disattivare il partizionamento senza dover modificare le impostazioni dei campi.
- **Suddivisioni.** Per i modelli di suddivisione, selezionare il campo o i campi di suddivisione. Questa operazione è simile all'impostazione del ruolo di un campo su *Suddivisione* in un nodo Tipo. È possibile designare come campi di suddivisione solo i campi con livello di misurazione **Flag**, **Nominale**, **Ordinale** o **Continuo**. I campi selezionati come campi di suddivisione non possono essere utilizzati come campi obiettivo, di input, di partizione, di frequenza o peso. Per ulteriori informazioni, consultare l'argomento “Creazione di modelli di suddivisione” a pagina 28.
- **Utilizza campo frequenza.** Questa opzione consente di selezionare un campo come ponderazione della frequenza. Utilizzarla se i record nei dati di addestramento rappresentano più di una unità ciascuno, per esempio se si utilizzano dati aggregati. I valori del campo dovrebbero essere rappresentati dal numero di unità rappresentate da ciascun record. Per ulteriori informazioni, consultare l'argomento “Utilizzo dei campi frequenza e peso.” a pagina 33.

Nota: se viene visualizzato il messaggio di errore **Metadati non validi (su campi di input/output)**, verificare che siano stati specificati tutti i campi richiesti, come, ad esempio, il campo frequenza.

- **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. Per ulteriori informazioni, consultare l'argomento “Utilizzo dei campi frequenza e peso.” a pagina 33.
- **Conseguenti.** Per i nodi induzione di regole (Apriori), selezionare i campi da utilizzare come conseguenti nell'insieme di regole risultante. Corrisponde a campi con ruolo *Obiettivo* o *Entrambi* in un nodo Tipo.

- **Antecedenti.** Per i nodi induzione di regole (Apriori), selezionare i campi da utilizzare come antecedenti nell'insieme di regole risultante. Corrisponde a campi con ruolo *Input* o *Entrambi* in un nodo Tipo.

Alcuni modelli hanno una scheda Campi che differisce da quelli descritti in questa sezione.

- Per ulteriori informazioni, consultare l'argomento "Opzioni dei campi del nodo Sequenza" a pagina 246.
- Per ulteriori informazioni, consultare l'argomento "Opzioni dei campi del nodo CARMA" a pagina 234.

Utilizzo dei campi frequenza e peso.

I campi frequenza e peso vengono utilizzati per conferire maggiore importanza ad alcuni record rispetto ad altri, per esempio quando si sa che una sezione della popolazione è sottorappresentata nei dati di addestramento (peso) o quando un record rappresenta diversi casi identici (frequenza).

- i valori per un campo frequenza devono essere numeri interi positivi. I record con frequenza negativa o zero sono esclusi dall'analisi. Le frequenze non intere sono arrotondate al numero intero più vicino.
- I valori dei pesi del caso devono essere positivi, ma non necessariamente dei numeri interi. I record con peso del caso negativo o pari a zero sono esclusi dall'analisi.

Calcolo del punteggio dei campi frequenza e peso

I campi frequenza e peso sono utilizzati nell'addestramento dei modelli ma non nel calcolo del punteggio, poiché il punteggio di ogni record si basa sulle sue caratteristiche, indipendentemente dal numero di casi che rappresenta. Ad esempio, si supponga che i dati siano contenuti nella seguente tabella.

Tabella 1. Esempio di dati

Coniugato	**Risposta**
Sì	Sì
Sì	Sì
Sì	Sì
Sì	No
No	Sì
No	No
No	No

In base a tali informazioni, tre delle quattro persone coniugate hanno risposto alla promozione e due delle tre persone non coniugate non hanno risposto. Il calcolo del punteggio dei nuovi record viene eseguito di conseguenza, come illustrato nella tabella riportata di seguito.

Tabella 2. Esempio di record di cui è stato calcolato il punteggio

Coniugato	**$-Risposta**	**$RP-Risposta**
Sì	Sì	0.75 (tre su quattro)
No	No	0.67 (due su tre)

In alternativa, è possibile archiviare i propri dati di addestramento in modo più compatto, utilizzando un campo frequenza, come illustrato nella tabella riportata di seguito.

Tabella 3. Esempio alternativo di record di cui è stato calcolato il punteggio

Coniugato	Risposta	Frequenza
Sì	Sì	3
Sì	No	1
No	Sì	1
No	No	2

Poiché l'insieme di dati rappresentato è esattamente lo stesso, verrà creato lo stesso modello e le risposte saranno previste solo in base allo stato civile. Se i dati per il calcolo del punteggio comprendono dieci soggetti coniugati, la previsione sarà *Sì* per ciascuno di essi, indipendentemente dal fatto che siano presentati come dieci record separati o come un singolo record con un valore di frequenza di 10. Il peso, benché non sia generalmente un numero intero, può essere considerato un indicatore analogo dell'importanza di un record. Questo è il motivo per cui i campi frequenza e peso non vengono utilizzati nel calcolo del punteggio dei record.

Valutazione e confronto di modelli

Alcuni tipi di modelli supportano i campi frequenza, altri supportano i campi peso e altri ancora li supportano entrambi. Tuttavia, in tutti i casi in cui sono validi, essi sono utilizzati solo per la creazione del modello e non vengono considerati nella valutazione dei modelli mediante un nodo Valutazione o Analisi né nella classificazione dei modelli eseguita tramite gran parte dei metodi supportati dai nodi Classificatore automatico e Numerico automatico.

- Per esempio, quando si confrontano modelli tramite grafici di valutazione, i valori di frequenza e peso vengono ignorati. Ciò consente il confronto di livelli tra i modelli che utilizzano questi campi ed i modelli che non li utilizzano, ma indica che, per una valutazione accurata, è necessario utilizzare un insieme di dati che rappresenta in modo preciso la popolazione senza fare affidamento su un campo frequenza o peso. In pratica, per ottenere questo risultato è necessario verificare che i modelli siano valutati mediante un campione di test in cui il valore del campo frequenza o peso sia sempre null o pari a 1 (questa limitazione si applica solo quando si valutano i modelli; se frequenza o peso fossero sempre pari a 1 sia per i campioni di addestramento che per quelli di test sarebbe del tutto inutile utilizzare questi campi).
- Se si utilizza il nodo Classificatore automatico è possibile tenere conto della frequenza se i modelli vengono classificati in base al Profitto; pertanto, in quel caso specifico, questo metodo è consigliato.
- Se necessario è possibile suddividere i dati in campioni di addestramento e campioni di test mediante un nodo Partizione.

Opzioni della scheda Analizza di un nodo Modelli

Molti nodi di modellazione dispongono di una scheda Analisi che consente di ottenere informazioni relative all'importanza predittore insieme ai punteggi di propensione grezza e corretta.

Valutazione del modello

Calcola importanza predittori. Per i modelli che generano una misura appropriata dell'importanza è possibile visualizzare un grafico che indica l'importanza relativa di ogni predittore nella stima del modello. In genere è opportuno concentrare la modellazione sui predittori più importanti e valutare se eliminare o ignorare quelli meno importanti. Notare che per calcolare l'importanza dei predittori per alcuni modelli potrebbe essere necessario più tempo, in particolare quando vengono utilizzati dataset di grandi dimensioni e che, come risultato, la funzione è disattivata per impostazione predefinita per alcuni modelli. L'importanza dei predittori non è disponibile per i modelli di elenco di decisioni. Per ulteriori informazioni, consultare l'argomento “Importanza predittore” a pagina 43.

Punteggi di propensione

I punteggi di propensione possono essere attivati nel nodo Modelli e nella scheda Impostazioni del nugget del modello. Questa funzionalità è disponibile solamente quando l'obiettivo selezionato è un campo flag. Per ulteriori informazioni, consultare l'argomento “Punteggi di propensione”.

Calcola punteggi di propensione grezza. I punteggi di propensione grezza vengono derivati dal modello in base ai soli dati di addestramento. Se il modello prevede il valore *vero* (risposta favorevole), la propensione sarà uguale a P, dove P è la probabilità della previsione. SE il modello prevede il valore falso, la propensione viene calcolata come (1 – P).

- Se si seleziona questa opzione durante la creazione del modello, nel nugget del modello i punteggi di propensione saranno attivati per default. Tuttavia, nel nugget del modello si può sempre decidere di attivare i punteggi di propensione grezza indipendentemente dal fatto che siano stati selezionati nel nodo Modelli.
- Quando si calcola il punteggio del modello, i punteggi di propensione grezza saranno aggiunti in un campo al cui prefisso standard vengono fatte seguire le lettere *RP*. Per esempio, se le previsioni sono in un campo denominato *$R-tasso di abbandono*, il nome del campo del punteggio di propensione sarà *$RRP-tasso di abbandono*.

Calcola punteggi di propensione regolata. Le propensioni grezze si basano unicamente sulle stime fornite dal modello, le quali possono essere sovradattate e determinare di conseguenza delle stime eccessivamente ottimistiche della propensione. Le propensioni regolate cercano di compensare esaminando le prestazioni del modello sulle partizioni di test e di convalida e adeguando le propensioni di conseguenza per fornire una stima più corretta.

- Questa impostazione richiede la presenza di un campo partizione valido nel flusso.
- Al contrario dei punteggi di confidenza grezza, i punteggi di propensione regolata devono essere calcolati in sede di creazione del modello, altrimenti non saranno disponibili quando si calcola il punteggio del nugget del modello.
- Quando si calcola il punteggio del modello, i punteggi di propensione regolata saranno aggiunti in un campo al cui prefisso standard vengono fatte seguire le lettere *AP*. Per esempio, se le previsioni sono in un campo denominato *$R-tasso di abbandono*, il nome del campo del punteggio di propensione sarà *$RAP-tasso di abbandono*. I punteggi di propensione regolata non sono disponibili per i modelli di regressione logistica.
- Quando si calcolano i punteggi di propensione regolata, la partizione di test o di convalida utilizzata per il calcolo non deve essere stata bilanciata. A tal fine, verificare che l'opzione **Bilancia solo dati di addestramento** sia selezionata in tutti i nodi bilanciamento a monte. Inoltre, se a monte è presente un campione complesso, i punteggi di propensione regolata saranno invalidati.
- I punteggi di propensione regolata non sono disponibili per i modelli di struttura ad albero boosted e di insiemi di regole. Per ulteriori informazioni, consultare l'argomento “Modelli C5.0 boosted” a pagina 113.

In base a. Per consentire il calcolo del punteggio di propensione regolata, nel flusso deve essere presente un campo partizione. È possibile specificare se per il calcolo deve essere utilizzata la partizione di test o di convalida. Per ottenere risultati migliori, la partizione di test o di convalida deve contenere un numero di record pari almeno a quelli presenti nella partizione utilizzata per l'addestramento del modello originale.

Punteggi di propensione

Per i modelli che restituiscono una previsione di tipo *sì* o *no* è possibile richiedere il calcolo del punteggio di propensione oltre ai normali valori di previsione e confidenza. I punteggi di propensione indicano la probabilità di un determinato risultato o risposta. La tabella riportata di seguito contiene un esempio.

Tabella 4. Punteggi di propensione

Cliente	Propensione alla risposta
Giovanni Bianchi	35%
Giovanna Bianchi	15%

I punteggi di propensione sono disponibili solo per i modelli con obiettivi flag e indicano la probabilità del valore *Vero* definito per il campo, come specificato in un nodo origine o Tipo.

Punteggi di propensione e punteggi di confidenza

I punteggi di propensione sono diversi dai punteggi di confidenza, che vengono applicati alla previsione corrente, sia per i casi in cui il valore è *sì* sia per quelli in cui il valore è *no*. Nei casi in cui la previsione è *no*, ad esempio, una confidenza elevata indica una probabilità elevata di *non* rispondere. I punteggi di propensione sfuggono a questa limitazione per consentire un più semplice confronto tra tutti i record. Per esempio, una previsione *no* con una confidenza di *0.85* si traduce in una propensione grezza di *0.15* (o *1 meno 0.85*).

Tabella 5. Punteggi di confidenza

Cliente	Previsione	Confidenza
Giovanni Bianchi	Risposta	.35
Giovanna Bianchi	Mancata risposta	.85

calcolo del punteggio di propensione

- I punteggi di propensione possono essere attivati nella scheda Analisi del nodo Modelli o nella scheda Impostazioni del nugget del modello. Questa funzionalità è disponibile solamente quando l'obiettivo selezionato è un campo flag. Per ulteriori informazioni, consultare l'argomento "Opzioni della scheda Analizza di un nodo Modelli" a pagina 34.
- A seconda del metodo dell'insieme utilizzato, i punteggi di propensione si possono calcolare anche mediante il nodo dell'insieme.

Calcolo dei punteggi di propensione regolata

I punteggi di propensione regolata vengono calcolati in sede di creazione del modello e non sono disponibili in altri casi. Una volta creato il modello, ne viene calcolato il punteggio utilizzando i dati dalla partizione di convalida o di test e viene creato un nuovo modello per il calcolo dei punteggi di propensione regolata analizzando le prestazioni del modello originale su tale partizione. Per calcolare i punteggi di propensione regolata è possibile utilizzare due diversi metodi che dipendono dal tipo di modello.

- Per i modelli di struttura ad albero e Insieme di regole, i punteggi di propensione regolata vengono generati ricalcolando la frequenza delle singole categorie in corrispondenza di ogni nodo della struttura ad albero (per i modelli di struttura ad albero) o il supporto e la confidenza delle singole regole (per i modelli Insieme di regole). Il calcolo genera un nuovo modello di struttura ad albero o Insieme di regole che viene archiviato insieme al modello originale per essere utilizzato ogni volta che vengono richiesti punteggi di propensione regolata. Ogni volta che il modello originale viene applicato a dati nuovi, il nuovo modello può in seguito essere applicato ai punteggi di propensione grezza per generare i punteggi regolati.
- Per gli altri modelli, i record generati calcolando il punteggio del modello originale sulla partizione di convalida o di test vengono in seguito discretizzati in base ai rispettivi punteggi di propensione grezza. Viene quindi addestrato un modello Rete neurale che definisce una funzione non lineare che mappa dalla propensione grezza media di ogni bin alla propensione osservata media dello stesso bin. Come per i modelli di struttura ad albero citati sopra citati, il modello Rete neurale risultante viene archiviato

insieme al modello originale e può essere applicato ai punteggi di propensione grezza tutte le volte che sono necessari dei punteggi di propensione regolata.

Avvertenza in merito ai valori mancanti nella partizione di test. La gestione dei valori di input mancanti nella partizione di test/convalida varia in base al modello (per i dettagli, consultare i singoli algoritmi di calcolo del punteggio del modello). Il modello C5 non può calcolare le propensioni regolate quando mancano dei valori di input.

Nugget del modello

Figura 19. Nugget del modello

Un nugget del modello è un contenitore per un modello, ovvero per l'insieme di regole, formule o equazioni che rappresentano i risultati del processo di creazione del modello in IBM SPSS Modeler. Lo scopo principale di un nugget è il calcolo del punteggio dei dati per generare previsioni o consentire un'ulteriore analisi delle proprietà del modello. Quando si apre un nugget del modello sullo schermo è possibile visualizzare una serie di dettagli del modello, per esempio l'importanza relativa dei campi di input nella creazione del modello stesso. Per visualizzare le previsioni è necessario collegare ed eseguire un ulteriore processo o nodo di output. Per ulteriori informazioni, consultare l'argomento "Utilizzo dei nugget del modello nei flussi" a pagina 48.

Figura 20. Collegamento del modello dal nodo Modelli al nugget del modello

Quando si esegue un nodo Modelli, nell'area del flusso viene inserito un nugget del modello corrispondente, rappresentato da un'icona che raffigura un diamante dorato. Nell'area del flusso, il nugget è visualizzato con una connessione (linea continua) al nodo appropriato più vicino a monte del nodo Modelli e un collegamento (linea tratteggiata) al nodo Modelli stesso.

Il nugget viene inoltre inserito nella palette Modelli nell'angolo superiore destro della finestra di IBM SPSS Modeler. I nugget possono essere selezionati da una di queste posizioni e consultati per visualizzare i dettagli del modello.

I nugget vengono sempre inseriti nella palette Modelli una volta conclusa l'esecuzione di un nodo Modelli. È possibile impostare un'opzione utente per controllare se il nugget viene collocato anche nell'area del flusso.

Gli argomenti seguenti forniscono informazioni sull'utilizzo dei nugget del modello in IBM SPSS Modeler. Per un'analisi approfondita degli algoritmi utilizzati, vedere *IBM SPSS ModelerAlgorithms Guide*, disponibile nella cartella *\Documentation* del DVD di IBM SPSS Modeler.

Collegamenti dei modelli

Per default, nell'area un nugget viene visualizzato con un collegamento al nodo Modelli a partire da cui è stato creato. Questo è utile soprattutto in presenza di flussi complessi formati da più nugget poiché permette di individuare l'insieme che sarà aggiornato da ogni nodo Modelli. Ogni collegamento contiene

un simbolo indicante se il modello viene sostituito all'esecuzione del nodo Modelli. Per ulteriori informazioni, consultare l'argomento “Sostituzione di un modello” a pagina 39.

Definizione ed eliminazione dei collegamenti dei modelli

I collegamenti si possono definire ed eliminare manualmente nell'area. Quando si definisce un nuovo collegamento, il cursore si trasforma in cursore del collegamento.

Figura 21. Cursore del collegamento

Definizione di un nuovo collegamento (menu di scelta rapida)

1. Con il pulsante destro del mouse, fare clic sul nodo Modelli da cui dovrà partire il collegamento.
2. Scegliere **Definisci collegamento del modello** dal menu di scelta rapida.
3. Fare clic sul nugget dove deve terminare il collegamento.

Definizione di un nuovo collegamento (menu principale)

1. Fare clic sul nodo Modelli da cui dovrà partire il collegamento.
2. Dal menu principale, scegliere:
   Modifica > **Nodo** > **Definisci collegamento del modello**
3. Fare clic sul nugget dove deve terminare il collegamento.

Eliminazione di un collegamento esistente (menu di scelta rapida)

1. Fare clic con il pulsante destro del mouse sul nugget a un'estremità del collegamento.
2. Scegliere **Rimuovi collegamento del modello** dal menu di scelta rapida.

In alternativa:

1. Fare clic con il pulsante destro del mouse sul simbolo al centro del collegamento.
2. Scegliere **Rimuovi collegamento** dal menu di scelta rapida.

Eliminazione di un collegamento esistente (menu principale)

1. Fare clic sul nodo Modelli o sul nugget da cui si desidera eliminare il collegamento.
2. Dal menu principale, scegliere:
   Modifica > **Nodo** > **Rimuovi collegamento del modello**

Copiare e incollare i collegamenti dei modelli

Se si copia un nugget dotato di collegamento senza il relativo nodo Modelli e lo si incolla nello stesso flusso, il nugget viene incollato con un collegamento al nodo Modelli. Il nuovo collegamento ha lo stesso stato di sostituzione del modello (consultare “Sostituzione di un modello” a pagina 39) del collegamento originale.

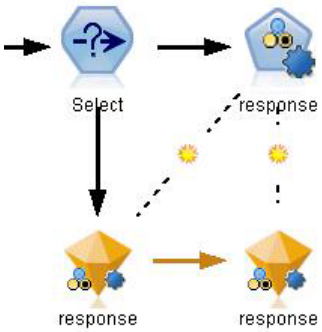


Figura 22. Copiare e incollare un nugget dotato di collegamento

Se si copia e incolla un nugget con il relativo nodo di modellazione collegato, il collegamento viene conservato indipendentemente dal fatto che gli oggetti vengano incollati nello stesso flusso o in un nuovo flusso.

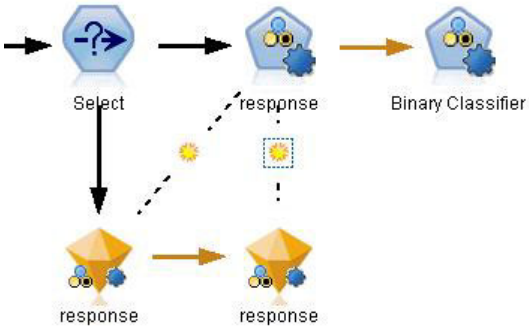


Figura 23. Copiare e incollare un nugget dotato di collegamento

Nota: se un nugget dotato di collegamento, senza il relativo nodo di modellazione, viene copiato ed incollato in un nuovo flusso (o in un Supernodo che non contiene il nodo di modellazione), il collegamento viene interrotto e viene copiato solo il nugget.

Collegamenti dei modelli e Supernodi

Se si definisce un Supernodo in modo da includere il nodo Modelli o il nugget del modello di un modello dotato di collegamento (ma non entrambi), il collegamento viene interrotto. Se si espande il Supernodo, il collegamento non viene ripristinato; questo è possibile solo annullando la creazione del Supernodo.

Sostituzione di un modello

È possibile decidere di sostituire (cioè aggiornare) un nugget esistente quando viene rieseguito il nodo Modelli a partire da cui l'insieme è stato creato. Se si disattiva l'opzione di sostituzione, quando si riesegue il nodo Modelli viene creato un nuovo nugget.

Nota: la sostituzione di un modello è diversa dall'aggiornamento di un modello, che fa riferimento all'aggiornamento di un modello in uno scenario.

Ogni collegamento tra un nodo Modelli e un nugget contiene un simbolo indicante se il modello viene sostituito a ogni nuova esecuzione del nodo Modelli.

Figura 24. Collegamento del modello con sostituzione del modello attivata

All'inizio, il collegamento è visualizzato con la sostituzione del modello attivata, rappresentata dal simboletto del raggio di sole. In queste condizioni, la riesecuzione del nodo Modelli a un'estremità del collegamento determina semplicemente l'aggiornamento del nugget all'altra estremità.

Figura 25. Collegamento del modello con sostituzione del modello disattivata

Se la sostituzione del modello è disattivata, il simbolo del collegamento è sostituito da un punto grigio. In queste condizioni, la riesecuzione del nodo Modelli a un'estremità del collegamento determina la creazione di una nuova versione aggiornata del nugget nell'area.

In entrambi i casi, nella palette Modelli viene aggiornato il nugget esistente o viene aggiunto un nuovo nugget, a seconda dell'impostazione dell'opzione di sistema **Sostituisci modello precedente**.

Ordine di esecuzione

Quando si esegue un flusso con più rami contenenti nugget del modello, il flusso viene prima valutato per assicurarsi che un ramo con sostituzione del modello attivata venga eseguito prima di tutti i rami che utilizzano il nugget del modello risultante.

Se i propri requisiti sono più complessi, è possibile impostare manualmente l'ordine di esecuzione tramite gli script.

Modifica dell'impostazione di sostituzione del modello

Per modificare l'impostazione di sostituzione del modello:

1. Fare clic con il pulsante destro del mouse sul simbolo presente nel collegamento.
2. Scegliere **Attiva (Disattiva) sostituzione del modello**, a seconda delle esigenze.

Nota: l'impostazione di sostituzione del modello su un collegamento del modello sostituisce l'impostazione nella scheda Notifiche della finestra di dialogo Opzioni utente (Strumenti > Opzioni > Opzioni utente).

La palette Modelli

La tavolozza dei modelli (nella scheda Modelli nella finestra Manager) consente di utilizzare, esaminare e modificare i nugget del modello in diversi modi.

Figura 26. Palette Modelli

Se si fa clic con il pulsante destro del mouse su un nugget del modello nella palette Modelli si apre un menu di scelta rapida con le opzioni seguenti:

- **Aggiungi a flusso.** Aggiunge il nugget del modello al flusso correntemente attivo. Se nel flusso viene selezionato un nodo, il nugget del modello verrà collegato al nodo selezionato quando tale collegamento è possibile, o altrimenti al nodo appropriato più vicino. Il nugget è visualizzato con un collegamento al nodo Modelli da cui è stato creato, se il nodo è ancora presente nel flusso.
- **Sfoglia.** Apre il browser del modello relativo al nugget.
- **Rinomina e annota.** Consente di rinominare il nugget del modello e/o di modificare l'annotazione relativa al nugget.
- **Genera nodo modellazione.** Se si desidera modificare o aggiornare un nugget del modello ma il flusso utilizzato per creare il modello non è più disponibile, è possibile utilizzare questa opzione per rigenerare un nodo Modelli con le stesse opzioni utilizzate per creare il modello originale.
- **Salva modello, Salva modello con nome.** Salva il nugget del modello in un file binario esterno modello generato (.gm).
- **Archivia modello.** Archivia il nugget del modello in IBM SPSS Collaboration and Deployment Services Repository.
- **Esporta PMML.** Esporta il nugget del modello come PMML (Predictive Model Markup Language), che può essere utilizzato per determinare il punteggio dei nuovi dati in programmi diversi da IBM SPSS Modeler. **Esporta PMML** è disponibile per tutti i nodi dei modelli generati. *Nota*: per utilizzare questa funzione, è necessaria una licenza per IBM SPSS Modeler Server.
- **Aggiungi a progetto.** Salva il nugget del modello e lo aggiunge al progetto corrente. Nella scheda Classi, il nugget del modello verrà aggiunto alla cartella Modelli generati. Nella scheda CRISP-DM, verrà aggiunto alla fase del progetto di default.
- **Elimina.** Elimina il nugget del modello dalla palette.

Se si fa clic con il pulsante destro del mouse su un'area libera della palette Modelli si apre un menu di scelta rapida con le opzioni seguenti:

- **Apri modello.** Carica un nugget del modello generato precedentemente in IBM SPSS Modeler.
- **Recupera modello.** Recupera un modello archiviato da un repository IBM SPSS Collaboration and Deployment Services.
- **Carica tavolozza.** Carica una palette Modelli salvata da un file esterno.
- **Recupera tavolozza.** Recupera una palette Modelli archiviata da un repository IBM SPSS Collaboration and Deployment Services.
- **Salva tavolozza.** Salva tutto il contenuto della palette Modelli in un file esterno palette Modelli generati (.gen).

- **Archivia tavolozza.** Memorizza tutto il contenuto della palette Modelli in un repository IBM SPSS Collaboration and Deployment Services.
- **Cancella tavolozza.** Elimina tutti i nugget dalla palette.
- **Aggiungi tavolozza a progetto.** Salva la palette Modelli e la aggiunge al progetto corrente. Nella scheda Classi, il nugget del modello verrà aggiunto alla cartella Modelli generati. Nella scheda CRISP-DM, verrà aggiunto alla fase del progetto di default.
- **Importa PMML.** Carica un modello da un file esterno. È possibile aprire, sfogliare e calcolare il punteggio dei modelli PMML creati da IBM SPSS Statistics o da altre applicazioni che supportano questo formato. Per ulteriori informazioni, consultare l'argomento "Importazione ed esportazione di modelli come PMML" a pagina 49.

Esplorazione dei nugget del modello

I browser del nugget del modello consentono di esaminare ed utilizzare i risultati dei propri modelli. Dal browser, è possibile salvare, stampare o esportare il modello generato, esaminare il riepilogo del modello e visualizzare o modificare le annotazioni del modello. In alcuni tipi di nugget del modello è anche possibile generare nuovi nodi, per esempio i nodi Filtro o Insieme di regole. In altri modelli, è anche possibile visualizzare i parametri del modello, quali regole o centri di cluster. Per alcuni tipi di modelli (modelli cluster e modelli basati su strutture ad albero), è possibile visualizzare una rappresentazione grafica della struttura del modello. Di seguito vengono descritti i comandi per l'utilizzo dei browser dei nugget del modello.

Menu

Menu File. In tutti i nugget del modello è presente un menu File che contiene un sottoinsieme delle opzioni seguenti:

- **Salva nodo.** Salva il nugget del modello in un file nodo (.nod).
- **Archivia nodo.** Archivia il nugget del modello in un repository IBM SPSS Collaboration and Deployment Services.
- **Intestazione e piè di pagina.** Consente di modificare l'intestazione e il piè di pagina della pagina per la stampa dal nugget del modello.
- **Imposta pagina.** Consente di modificare le impostazioni della pagina per la stampa dal nugget del modello.
- **Anteprima di stampa.** Visualizza un'anteprima di come il nugget del modello verrà visualizzato nella stampa. Selezionare le informazioni di cui si desidera vedere l'anteprima dal sottomenu.
- **Stampa.** Stampa il contenuto del nugget. Selezionare le informazioni che si desidera stampare dal sottomenu.
- **Stampa visualizzazione.** Stampa la visualizzazione corrente o tutte le visualizzazioni.
- **Esporta testo.** Esporta il contenuto del nugget del modello in un file di testo. Selezionare le informazioni che si desidera esportare dal sottomenu.
- **Esporta HTML.** Esporta il contenuto del nugget del modello in un file HTML. Selezionare le informazioni che si desidera esportare dal sottomenu.
- **Esporta PMML.** Esporta il modello come PMML (Predictive Model Markup Language), che può essere utilizzato con altri software compatibili con PMML. Per ulteriori informazioni, consultare l'argomento "Importazione ed esportazione di modelli come PMML" a pagina 49. *Nota*: per utilizzare questa funzione, è necessaria una licenza per IBM SPSS Modeler Server.
- **Esporta in SQL.** Esporta il modello sotto forma di codice SQL (Structured Query Language), che può essere modificato e utilizzato con altri database.

  Nota: l'esportazione in SQL è disponibile solo dai seguenti modelli: C5, C&RT, CHAID, QUEST, Regressione lineare, Regressione logistica, Rete neurale, Fattoriale/PCA ed Elenco di decisioni.
- **Pubblica per adattatore per calcolo del punteggio server.** Pubblica il modello in un database su cui è installato un adattatore per calcolo del punteggio, abilitando l'esecuzione del calcolo del punteggio del

modello all'interno del database. Per ulteriori informazioni, consultare l'argomento “Pubblicazione dei modelli per un adattatore per calcolo del punteggio” a pagina 51.

Menu Genera. La maggior parte dei nugget del modello dispongono anche di un menu Genera, che consente di generare nuovi nodi in base al nugget del modello. Le opzioni disponibili in questo menu dipendono dal tipo di modello che si sta visualizzando. Per informazioni sulle possibilità di generazione offerte da un determinato nugget del modello, vedere il tipo specifico di modello.

Menu Visualizza. Nella scheda Modello di un nugget, questo menu consente di visualizzare o nascondere le diverse barre degli strumenti di visualizzazione che sono disponibili nella modalità corrente. Per rendere disponibile l'insieme completo di barre degli strumenti, selezionare la modalità Modifica (l'icona a forma di pennello) dalla barra degli strumenti Generale.

Pulsante Anteprima. Alcuni nugget del modello sono dotati di un pulsante Anteprima che consente di visualizzare un campione di dati del modello, inclusi i campi addizionali creati dal processo di modellazione. Per impostazione predefinita, vengono visualizzate dieci righe. È tuttavia possibile modificare il valore nelle proprietà del flusso.

Pulsante Aggiungi a progetto corrente. Salva il nugget del modello e lo aggiunge al progetto corrente. Nella scheda Classi, il nugget del modello verrà aggiunto alla cartella Modelli generati. Nella scheda CRISP-DM, verrà aggiunto alla fase del progetto di default.

Scheda Riepilogo di un nugget del modello/Informazioni

La scheda Riepilogo o la visualizzazione Informazioni di un 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.

Analisi. Visualizza informazioni sul modello. I dettagli specifici variano a seconda dei tipi di modello e sono descritti nelle sezioni relative ai singoli nugget del modello. Inoltre, se è stato eseguito un nodo Analisi collegato a questo nodo Modelli, in questa sezione verranno visualizzate anche le informazioni sui risultati di tale analisi.

Campi. Elenca i campi utilizzati come obiettivi e come input nella creazione del modello. Per i modelli di suddivisione, elencare anche i campi che hanno determinato le suddivisioni.

Impostazioni/Opzioni di creazione. Contiene informazioni sulle impostazioni utilizzate nella creazione del modello.

Riepilogo addestramento. Mostra il tipo di modello, il flusso utilizzato per creare tale modello, l'utente che lo ha creato, la data di creazione e il tempo trascorso per la creazione del modello.

Importanza predittore

Generalmente, lo sforzo della modellazione viene concentrato sui campi predittore più importanti, senza considerare o ignorando i campi di minore importanza. Il grafico dell'importanza dei predittori rende più semplice questa operazione, indicando l'importanza relativa di ciascun predittore nella stima del modello. Poiché i valori sono relativi, la somma dei valori visualizzata per tutti i predittori è 1.0. L'importanza dei predittori non è correlata alla precisione del modello. Riguarda unicamente l'importanza di ciascun predittore nell'esecuzione di una previsione e non il grado di precisione della previsione.

L'importanza predittore è disponibile per i modelli che producono una misura statistica appropriata dell'importanza, come reti neurali, strutture ad albero delle decisioni (C&R Tree, C5.0, CHAID e QUEST), reti bayesiane, discriminanti, SVM e modelli SLRM, regressione lineare e logistica, modelli lineari

generalizzati e modelli KNN (elemento adiacente più vicino). Per quasi tutti questi modelli, l'importanza dei predittori può essere attivata nella scheda Analisi del nodo Modelli. Per ulteriori informazioni, consultare l'argomento “Opzioni della scheda Analizza di un nodo Modelli” a pagina 34. Per i modelli KNN, consultare “Elementi adiacenti” a pagina 289.

Nota: l'importanza predittore non è supportata per i modelli di suddivisione. Le impostazioni di importanza dei predittori vengono ignorate durante la creazione dei modelli di suddivisione. Per ulteriori informazioni, consultare l'argomento “Creazione di modelli di suddivisione” a pagina 28.

Il calcolo dell'importanza dei predittori può richiedere molto più tempo della creazione del modello, soprattutto quando si utilizzano insiemi di dati di grandi dimensioni. Rispetto agli altri modelli, tale calcolo risulta particolarmente lungo per i modelli SVM e di regressione logistica e in questi modelli è disattivato per default. Se si utilizza un insieme di dati con un gran numero di predittori, uno screening iniziale mediante il nodo Selezione funzioni può fornire risultati più rapidi (vedere di seguito).

- L'importanza dei predittori viene calcolata a partire dalla partizione di test, se disponibile; in caso contrario vengono utilizzati i dati di addestramento.
- Nei modelli SLRM è possibile determinare l'importanza dei predittori, ma il calcolo viene eseguito dall'algoritmo SLRM. Per ulteriori informazioni, consultare l'argomento “Nugget del modello SLRM” a pagina 278.
- Per interagire con il grafico, modificarlo e salvarlo è possibile utilizzare gli strumenti per i grafici di IBM SPSS Modeler,
- Se lo si desidera è possibile generare un nodo Filtro a partire dalle informazioni contenute nel grafico dell'importanza dei predittori. Per ulteriori informazioni, consultare l'argomento “Applicazione di filtri alle variabili a seconda dell'importanza” a pagina 45.

Importanza predittore e Selezione funzioni

In alcuni casi, il grafico dell'importanza dei predittori visualizzato in un nugget del modello sembra fornire risultati apparentemente simili a quelli del nodo Selezione funzioni. Mentre questa funzione classifica i singoli campi di input in base alla forza della loro relazione con l'obiettivo specificato, indipendentemente dagli altri input, il grafico dell'importanza dei predittori indica l'importanza relativa di ogni input per *quel* determinato modello. Pertanto, il nodo Selezione funzioni risulterà più conservativo nello screening degli input. Per esempio, se *posizione* e *categoria professionale* sono entrambe fortemente correlate allo stipendio, Selezione funzioni indicherebbe che entrambe sono importanti. Nella modellazione, invece, viene tenuto conto anche delle interazioni e delle correlazioni. Si potrebbe quindi notare che tra due input ne viene utilizzato solo uno se entrambi replicano in gran parte le stesse informazioni. In pratica, il nodo Selezione funzioni è utile soprattutto per lo screening preliminare, particolarmente per insiemi di dati di grandi dimensioni e con un gran numero di variabili, mentre l'importanza dei predittori i è più utile nell'ottimizzazione del modello.

Differenze dell'importanza predittore tra singoli modelli e nodi di modellazione automatizzata

A seconda della creazione di un unico modello da un nodo singolo o dall'utilizzo di un nodo di modellazione automatizzata per generare i risultati, si possono notare differenze minime nell'importanza predittore. Tali differenze nell'implementazione sono dovute a limitazioni tecniche.

Ad esempio, con singoli classificatori come il CHAID, il calcolo si applica alla regola di arresto e utilizza i valori di probabilità durante il calcolo dei valori di importanza. Invece, il classificatore automatico non utilizza una regola di arresto e utilizza le etichette previste direttamente nel calcolo. Tali differenze possono indicare che se si genera un singolo modello mediante il classificatore automatico, il valore di importanza può essere considerato una stima approssimativa in confronto al valore calcolato per un singolo classificatore. Per ottenere i valori di importanza predittore più accurati, si consiglia di utilizzare un singolo nodo anziché i nodi di modellazione automatizzata.

Applicazione di filtri alle variabili a seconda dell'importanza

Se lo si desidera è possibile generare un nodo Filtro a partire dalle informazioni contenute nel grafico dell'importanza dei predittori.

Contrassegnare gli eventuali predittori da includere nel grafico e dai menu selezionare:

Genera > **Nodo Filtro (Importanza predittore)**

OPPURE

> **Selezione campi (Importanza predittore)**

Prime N variabili. Include o esclude i predittori più importanti fino al numero specificato.

Importanza maggiore di. Include o esclude tutti i predittori con importanza relativa maggiore del valore specificato.

Visualizzatore di insieme

Modelli per insiemi

Il modello per un insieme fornisce informazioni relative ai modelli di componenti nell'insieme ed alle prestazioni dell'intero insieme.

La barra degli strumenti principale (indipendente dalla vista) consente di scegliere se utilizzare l'insieme o un modello di riferimento per il calcolo del punteggio. Se per il calcolo del punteggio viene utilizzato l'insieme, è anche possibile selezionare la regola di combinazione. Tali modifiche non richiedono di eseguire nuovamente il modello; tuttavia, le scelte vengono salvate nel modello (nugget)per il calcolo del punteggio e/o la valutazione del modello a valle. Inoltre, hanno effetto sul PMML esportato dal visualizzatore di insieme.

Regola di combinazione. Quando viene calcolato il punteggio di un insieme, questa è la regola utilizzata per combinare i valori previsti dai modelli di base per il calcolo del valore del punteggio dell'insieme.

- I valori previsti dell'insieme per i target **di categoria** possono essere combinati utilizzando la votazione, la probabilità più elevata o la probabilità media più elevata. La **votazione** consente di selezionare la categoria che presenta più spesso la probabilità più elevata nei modelli di base. La **probabilità più elevata** consente di selezionare la categoria che raggiunge la singola probabilità più elevata in tutti i modelli di base. La **probabilità media più elevata** consente di selezionare la categoria con il massimo valore quando viene calcolata la media delle probabilità della categoria in tutti i modelli di base.
- I valori previsti dell'insieme per i target **continui** possono essere combinati utilizzando la media o la mediana dei valori previsti dai modelli di base.

Il valore predefinito viene rilevato dalle specifiche eseguite durante la creazione del modello. Modificando la regola di combinazione, l'accuratezza del modello viene nuovamente calcolata e tutte le viste di precisione del modello vengono aggiornate. Inoltre, viene aggiornato anche il grafico Importanza predittore. Questo controllo è disabilitato se per il calcolo del punteggio viene selezionato il modello di riferimento.

Mostra tutte le regole di combinazione. Quando questa opzione è selezionata, nel grafico di qualità del modello vengono visualizzati i risultati per tutte le regole di combinazione disponibili. Viene aggiornato anche il grafico Precisione del modello di componente in modo da mostrare le linee di riferimento per ciascun metodo di votazione.

Riepilogo del modello: La vista Riepilogo del modello è una snapshot dell'insieme e ne visualizza la qualità e la diversità.

Qualità. Il grafico visualizza l'accuratezza del modello finale in confronto ad un modello di riferimento e ad un modello nativo. L'accuratezza viene visualizzata in formato "larger is better"; il modello migliore avrà l'accuratezza più alta. Per un target di categoria, l'accuratezza è semplicemente la percentuale di record per cui il valore previsto corrisponde al valore osservato. Per un target continuo, l'accuratezza è 1 meno il rapporto tra l'errore assoluto della media nella previsione (la media dei valori assoluti dei valori previsti meno i valori osservati) e l'intervallo dei valori previsti (il massimo valore previsto meno il minimo valore previsto).

Per gli insiemi di bagging, il modello di riferimento è un modello standard creato sull'intera partizione di addestramento. Per gli insiemi boosted, il modello di riferimento è il primo modello di componenti.

Il modello naive rappresenta l'accuratezza se non è stato creato alcun modello ed assegna tutti i record alla categoria modale. Il modello naive non viene calcolato per i target continui.

Diversità. Il grafico visualizza la "diversità di opinione" tra i modelli di componenti utilizzati per creare l'insieme, presentato in formato "larger is more diverse". Si tratta di una misura che indica la quantità di variazione delle previsioni tra i modelli di base. La diversità non è disponibile per i modelli dell'insieme boosted e non viene visualizzata per i target continui.

Importanza predittore: Generalmente, lo sforzo della modellazione viene concentrato sui campi predittore più importanti, senza considerare o ignorando i campi di minore importanza. Il grafico dell'importanza dei predittori rende più semplice questa operazione, indicando l'importanza relativa di ciascun predittore nella stima del modello. Poiché i valori sono relativi, la somma dei valori visualizzata per tutti i predittori è 1.0. L'importanza dei predittori non è correlata alla precisione del modello. Riguarda unicamente l'importanza di ciascun predittore nell'esecuzione di una previsione e non il grado di precisione della previsione.

L'importanza del predittore non è disponibile per tutti i modelli dell'insieme. La serie di predittori può variare tra i modelli di componenti, ma l'importanza può essere calcolata per i predittori utilizzati in almeno un modello di componente.

Frequenza dei predittori: L'insieme di predittori può variare tra i modelli di componenti in base alla scelta del metodo di modellazione o della selezione del predittore. Il grafico Frequenza dei predittori è un grafico a punti che mostra la distribuzione dei predittori tra i modelli di componenti nell'insieme. Ciascun punto rappresenta uno o più modelli di componenti che contengono il predittore. I predittori vengono tracciati sull'asse y e sono ordinati in ordine di frequenza decrescente; quindi, il predittore più in alto è quello utilizzato nel maggior numero di modelli di componenti, mentre quello più in basso è il predittore utilizzato nel minor numero di componenti. Sono visualizzati i primi 10 predittori.

I predittori visualizzati con maggior frequenza sono generalmente quelli più importanti. Questo grafico non è utile per i metodi in cui l'insieme di predittori non può variare tra i modelli di componenti.

Accuratezza del modello di componente: Questo grafico è un grafico a punti dell'accuratezza predittiva per i modelli del componente. Ciascun punto rappresenta uno o più modelli di componenti con il livello di accuratezza tracciato sull'asse y. Spostare il puntatore del mouse su un qualsiasi punto per visualizzare informazioni relative al singolo modello di componenti corrispondente.

Linee di riferimento. Il grafico visualizza linee colorate per l'insieme e per i modelli naïve ed il modello di riferimento. Accanto alla linea corrispondente al modello che verrà utilizzato per il calcolo del punteggio, viene visualizzato un segno di spunta.

Interattività. Il grafico viene aggiornato se viene modificata la regola di combinazione.

Insiemi boosted. Per gli insiemi boosted, viene visualizzato un grafico a linee.

Dettagli del modello di componenti: La tabella visualizza le informazioni relative ai modelli di componenti, elencati per riga. Per impostazione predefinita, i modelli di componenti sono ordinati in base all'ordine del numero di modello crescente. È possibile ordinare le righe in ordine crescente o decrescente in base ai valori di qualsiasi colonna.

Modello. Un numero che rappresenta l'ordine sequenziale in cui è stato creato il modello di componenti.

Precisione. L'accuratezza generale formattata come percentuale.

Metodo. Il metodo di modellazione.

Predittori. Il numero di predittori utilizzati nel modello di componenti.

Dimensione del modello. La dimensione del modello dipende dal metodo di modellazione: per le strutture ad albero, è il numero di nodi nella struttura ad albero; per i modelli lineari, è il numero di coefficienti; per le reti neurali è il numero di sinapsi.

Record. Il numero pesato di record di input nel campione di addestramento.

Preparazione automatica dati:

Questa visualizzazione contiene informazioni sui campi esclusi e sulla modalità di derivazione dei campi trasformati nel passaggio di preparazione automatica dei dati (ADP). Per ogni campo trasformato o escluso, la tabella indica il nome del campo, il ruolo nell'analisi e l'azione intrapresa nel passaggio dell'ADP. I campi sono ordinati in ordine alfabetico crescente in base al nome.

L'azione **Taglia valori anomali**, se visualizzata, indica che i valori dei predittori continui che superano un valore di interruzione (3 deviazioni standard rispetto alla media) sono stati impostati sul valore di interruzione.

Nugget del modello per i modelli di suddivisione

Il nugget del modello per un modello di suddivisione fornisce l'accesso a tutti i modelli separati creati dalle suddivisioni.

Un nugget del modello di suddivisione contiene:

- un elenco di tutti i modelli di suddivisione creati, con un insieme di statistiche relative a ciascun modello
- informazioni sul modello complessivo

Dall'elenco di modelli di suddivisione è possibile aprire singoli modelli per esaminarli più attentamente.

Visualizzatore del modello di suddivisione

La scheda Modello elenca tutti i modelli contenuti nel nugget e fornisce vari tipi di statistiche sui modelli di suddivisione. Dispone di due moduli generali, in base al nodo di modellazione.

Ordina per. Utilizzare questo elenco per scegliere l'ordine secondo il quale vengono elencati i modelli. È possibile ordinare l'elenco in base ai valori di una delle colonne di visualizzazione, in ordine crescente o decrescente. In alternativa fare clic sull'intestazione di una colonna per ordinare l'elenco secondo quella colonna. L'impostazione di default è l'ordine decrescente dei valori di precisione complessiva.

Mostra/nascondi menu di colonne. Fare clic su questo pulsante per visualizzare un menu da cui scegliere singole colonne da mostrare o nascondere.

Visualizza. Se si sta utilizzando il partizionamento, è possibile decidere se visualizzare i risultati per i dati di addestramento o per i dati di test.

Per ciascuna suddivisione i dettagli mostrati sono i seguenti:

Grafico. Una miniatura che indica la distribuzione dei dati per questo modello. Quando il nugget si trova nell'area, fare doppio clic sulla miniatura per aprire il grafico nelle dimensioni normali.

Modello. Un'icona del tipo di modello. Fare doppio clic sull'icona per aprire il nugget del modello per questa particolare suddivisione.

Campi di suddivisione. I campi designati nel nodo Modelli come campi di suddivisione, con i loro possibili valori.

N. record nella suddivisione. Il numero di record coinvolti in questa suddivisione particolare.

N. di campi utilizzati. Classifica i modelli di suddivisione in base al numero di campi di input utilizzati.

Precisione globale (%). La percentuale di record che viene prevista correttamente dal modello di suddivisione rispetto al numero totale di record in quella suddivisione.

Suddivisione. L'intestazione di colonna indica il campo o i campi utilizzati per la creazione delle suddivisioni, mentre le celle rappresentano i valori di suddivisione. Fare doppio clic su una qualsiasi suddivisione per aprire un visualizzatore del modello per il modello creato per tale suddivisione.

Precisione. L'accuratezza generale formattata come percentuale.

Dimensione del modello. La dimensione del modello dipende dal metodo di modellazione: per le strutture ad albero, è il numero di nodi nella struttura ad albero; per i modelli lineari, è il numero di coefficienti; per le reti neurali è il numero di sinapsi.

Record. Il numero pesato di record di input nel campione di addestramento.

Utilizzo dei nugget del modello nei flussi

È possibile collocare i nugget del modello nei flussi per determinare il punteggio dei nuovi dati e generare nuovi nodi. Il **calcolo del punteggio** dei dati consente di utilizzare le informazioni raccolte dalla creazione del modello per creare previsioni per i nuovi record. Per visualizzare i risultati del calcolo del punteggio è necessario collegare al nugget un nodo terminale (cioè un nodo di elaborazione o di output) ed eseguirlo.

Per alcuni modelli, i nugget del modello contengono informazioni aggiuntive sulla qualità della previsione, come i valori di confidenza o le distanze dai centri di cluster. La generazione di nuovi nodi consente di creare facilmente nuovi nodi in base alla struttura del modello generato. Ad esempio, la maggior parte dei modelli che eseguono la selezione dei campi di input consente di generare nodi Filtro che passano solo i campi di input identificati come importanti dal modello.

Per utilizzare un nugget del modello per il calcolo del punteggio dei dati

1. Collegare il nugget del modello a una sorgente dati o flusso che passi i dati al nodo.
2. Aggiungere o collegare uno o più nodi di output o di elaborazione (per esempio un nodo Tabella o un nodo Analisi) al nugget del modello.
3. Eseguire uno dei nodi downstream del nugget del modello.

Nota: non è possibile utilizzare il nodo Regola grezza per calcolare il punteggio dei dati. Per calcolare il punteggio dei dati in base ad un modello della regola di associazione, utilizzare il nodo Regola grezza

per generare un nugget Insieme di regole ed utilizzare tale nugget per il calcolo del punteggio. Per ulteriori informazioni, consultare l'argomento “Generazione di un insieme di regole da un nugget del modello di associazione” a pagina 241.

Per utilizzare un nugget del modello per la generazione di nodi di elaborazione

1. Visualizzare il modello nella palette o modificarlo nell'area del flusso.
2. Selezionare il tipo di nodo desiderato dal menu Genera della finestra del browser del nugget del modello. Le opzioni disponibili varieranno in base al tipo di nugget del modello. Per informazioni sulle possibilità di generazione offerte da un determinato nugget del modello, vedere il tipo specifico di modello.

Rigenerazione di un nodo Modelli

Se si desidera modificare o aggiornare un nugget del modello ma il flusso utilizzato per creare il modello non è più disponibile, è possibile rigenerare un nodo Modelli con le stesse opzioni utilizzate per creare il modello originale.

Per ricreare un modello, fare clic con il pulsante destro del mouse sul modello nella palette Modelli e selezionare **Genera nodo modellazione**.

In alternativa, durante la visualizzazione di un modello, scegliere **Genera nodo modellazione** dal menu Genera.

Il nodo Modelli rigenerato dovrebbe essere funzionalmente identico a quello utilizzato per creare il modello originale, nella maggior parte dei casi.

- Nel caso dei modelli di struttura ad albero delle decisioni, è possibile che con il nodo siano state archiviate ulteriori impostazioni specificate durante la sessione interattiva e l'opzione **Utilizza direttive della struttura ad albero** risulterà attivata nel nodo Modelli rigenerato.
- Nel caso dei modelli Elenco di decisioni, sarà attivata l'opzione **Utilizza informazioni sulla sessione interattiva salvata**. Per ulteriori informazioni, consultare l'argomento “Opzioni del modello Elenco di decisioni” a pagina 142.
- Per i modelli di serie temporali, l'opzione **Continua la stima mediante modelli esistenti** è abilitata; ciò consente di rigenerare il modello precedente con i dati correnti. Per ulteriori informazioni, consultare l'argomento “Opzioni del modello di serie temporali” a pagina 263.

Importazione ed esportazione di modelli come PMML

PMML, acronimo di Predictive Model Markup Language, è un formato XML per descrivere modelli statistici e di data mining, inclusi input ai modelli, trasformazioni utilizzate per la preparazione dei dati per il data mining e parametri che definiscono i modelli stessi. In IBM SPSS Modeler è possibile importare ed esportare PMML, rendendo possibile la condivisione di modelli con altre applicazioni che supportano questo formato, quali IBM SPSS Statistics.

Per ulteriori informazioni sul formato PMML, vedere il sito Web del gruppo di data mining all'indirizzo *http://www.dmg.org*.

Per esportare un modello:

L'esportazione in formato PMML è supportata per la maggior parte dei tipi di modelli generati in IBM SPSS Modeler. Consultare l'argomento “Tipi di modello che supportano il formato PMML” a pagina 50 per ulteriori informazioni.

1. Fare clic con il pulsante destro del mouse su un nugget del modello nella palette Modelli. In alternativa, fare doppio clic su un nugget del modello nell'area e selezionare il menu File.
2. Nel menu, fare clic su **Esporta PMML**.

3. Nella finestra di dialogo Esporta (o Salva), specificare una directory di destinazione e un nome univoco per il modello.

Nota: è possibile modificare opzioni per l'esportazione in formato PMML nella finestra di dialogo Opzioni utente. Nel menu principale, fare clic su:

Strumenti > **Opzioni** > **Opzioni utente**

e selezionare la scheda PMML.

Per importare un modello salvato come PMML

I modelli esportati come PMML da IBM SPSS Modeler o un'altra applicazione possono essere importati nella palette Modelli. Consultare l'argomento “Tipi di modello che supportano il formato PMML” per ulteriori informazioni.

1. Nella palette Modelli, fare clic con il pulsante destro del mouse sulla palette e selezionare **Importa PMML** dal menu.
2. Selezionare il file da importare e specificare le opzioni per le etichette di variabile in base alle proprie esigenze.
3. Fare clic su **Apri**.

Utilizza etichette variabili se presenti nel modello. Nel formato PMML è possibile specificare nomi ed etichette di variabile (per esempio ID referente per *RefID*) per le variabili contenute nel dizionario dati. Selezionare questa opzione per utilizzare le etichette di variabile, se presenti nel PMML esportato.

Se sono state selezionate le opzioni delle etichette di variabile ma non sono presenti etichette di variabile nel PMML, i nomi delle variabili verranno utilizzati normalmente.

Tipi di modello che supportano il formato PMML

Esportazione PMML

Modelli IBM SPSS Modeler. I seguenti modelli creati in IBM SPSS Modeler possono essere esportati come PMML 4.0:

- C&R Tree
- QUEST
- CHAID
- Regressione lineare
- Rete neurale
- C5.0
- Regressione logistica
- Genlin
- SVM
- Apriori
- Carma
- Medie K
- Kohonen
- TwoStep
- GLMM (il supporto è solo per modelli GLMM solo ad effetti fissi)
- Elenco di decisioni
- Cox
- Sequenza (calcolo del punteggio per i modelli PMML Sequenza non supportato)

- Modello Statistics

Modelli nativi di database. Per i modelli generati utilizzando algoritmi nativi di database, l'esportazione PMML è disponibile solo per i modelli IBM InfoSphere Warehouse. I modelli creati utilizzando Analysis Services di Microsoft o Oracle Data Miner non possono essere esportati. Si noti inoltre che i modelli IBM esportati come PMML non possono essere reimportati in IBM SPSS Modeler.

Importazione PMML

IBM SPSS Modeler può importare e calcolare il punteggio di modelli PMML generati dalle versioni correnti di tutti i prodotti IBM SPSS Statistics, compresi i modelli esportati da IBM SPSS Modeler nonché i modelli o le trasformazioni PMML generate da IBM SPSS Statistics 17.0 o versioni successive. In sostanza, sono importabili tutti i modelli PMML di cui il modulo di gestione del punteggio è in grado di calcolare il punteggio, con le seguenti eccezioni:

- Non è possibile importare modelli Apriori, CARMA, Rilevamento anomalie e Sequenza.
- Non è possibile sfogliare i modelli PMML dopo averli importati in IBM SPSS Modeler, anche se è possibile utilizzarli per il calcolo del punteggio. Questa eccezione riguarda anche i modelli che sono stati originariamente esportati da IBM SPSS Modeler. Per evitare tale limitazione, esportare il modello come file di modello generato (*.*gm*) anziché come PMML.
- I modelli IBM InfoSphere Warehouse esportati come PMML non possono essere importati.
- Si è verificata una convalida limitata all'importazione, ma la convalida completa viene eseguita al tentativo di calcolare il punteggio del modello. Pertanto è possibile che l'importazione riesca e che, invece, il calcolo del punteggio fallisca o produca risultati non corretti.

Pubblicazione dei modelli per un adattatore per calcolo del punteggio

È possibile pubblicare i modelli su un server di database su cui è installato un adattatore per calcolo del punteggio. Un adattatore per calcolo del punteggio consente di eseguire il calcolo del punteggio del modello all'interno del database utilizzando le funzionalità UDF (user-defined function - funzione definita dall'utente) del database. L'esecuzione del calcolo del punteggio nel database evita la necessità di estrarre i dati prima di calcolare il punteggio. La pubblicazione su un adattatore per calcolo del punteggio genera anche del codice SQL di esempio per eseguire l'UDF.

Per eseguire la pubblicazione su un adattatore per calcolo del punteggio

1. Fare doppio clic sul nugget del modello per aprirlo.
2. Dal menu del nugget del modello, selezionare:
   File > **Pubblica per Adattatore per calcolo del punteggio server**
3. Compilare i campi rilevanti nella finestra di dialogo e fare clic su **OK**.

Connessione database. I dettagli di connessione per il database che si desidera utilizzare per il modello.

ID pubblicazione. (Solo database DB2 per z/OS). Se viene ricreato lo stesso modello e viene utilizzato lo stesso ID pubblicazione, il codice SQL generato non viene modificato, per cui è possibile ricreare un modello senza che sia necessario modificare l'applicazione che utilizza il codice SQL precedentemente generato. Per altri database, il codice SQL generato è univoco per il modello.

Genera SQL di esempio. Se questa opzione è selezionata, il codice SQL di esempio viene generato all'interno del file specificato nel campo **File**.

Modelli grezzi

Un modello grezzo contiene informazioni estratte dai dati ma non è progettato per generare direttamente le previsioni. Pertanto, non può essere aggiunto ai flussi. I modelli grezzi sono visualizzati come "diamanti grezzi" nella tavolozza dei modelli generati.

Figura 27. Icona di modello grezzo

Per visualizzare le informazioni relative al modello Regola grezza, fare clic con il pulsante destro del mouse sul modello e scegliere **Visualizza** dal menu di scelta rapida. Come per gli altri modelli generati in IBM SPSS Modeler, le diverse schede forniscono informazioni di riepilogo e sulle regole relative al modello creato.

Generazione di nodi. Il menu Genera consente di creare nuovi nodi basati sulle regole.

- **Nodo Seleziona.** Genera un nodo Seleziona per selezionare i record a cui si applica la regola selezionata correntemente. Questa opzione non è attiva se non viene selezionata una regola.
- **Insieme di regole.** Genera un nodo Insieme di regole per prevedere i valori relativi a un campo obiettivo singolo. Per ulteriori informazioni, consultare l'argomento "Generazione di un insieme di regole da un nugget del modello di associazione" a pagina 241.

Capitolo 4. Modelli di screening

Screening di campi e record

Nelle fasi preliminari di un'analisi è possibile utilizzare numerosi nodi Modelli per individuare i campi e i record che probabilmente saranno di maggior interesse ai fini della modellazione. È possibile utilizzare il nodo Selezione funzioni per sottoporre a screening e classificare i campi per importanza e il nodo Rilevamento anomalie per individuare i record insoliti che non sono conformi agli schemi noti di dati "normali".

Il nodo Selezione funzioni effettua lo screening dei campi di input, rimuovendoli in base a un insieme di criteri quali la percentuale di valori mancanti. Classifica quindi gli input restanti in ordine di importanza rispetto a un determinato obiettivo. Per esempio, dato un insieme di dati con centinaia di input potenziali, quali sono quelli con la maggiore probabilità di essere utili nella modellazione di risultati clinici?

Il nodo Rilevamento anomalie identifica casi insoliti, o valori anomali, non conformi a schemi di dati “normali”. Con questo nodo è possibile identificare valori anomali anche se questi non rientrano in schemi precedentemente conosciuti e anche se l'utente non sa esattamente ciò che sta cercando.

Si noti che il rilevamento delle anomalie individua i record o i casi insoliti attraverso l'analisi dei cluster basata sull'insieme di campi selezionati nel modello, senza considerare alcun campo obiettivo (dipendente) specifico e indipendentemente dal fatto che questi campi siano pertinenti allo schema che si sta tentando di prevedere. Per questo motivo, l'utente potrebbe voler utilizzare il rilevamento di anomalie in combinazione con la selezione delle funzioni o un'altra tecnica per sottoporre a screening e classificare i campi. Per esempio, è possibile utilizzare la selezione delle funzioni per individuare i campi più importanti relativi a un obiettivo specifico e quindi utilizzare il rilevamento di anomalie per individuare i record più insoliti rispetto a tali campi (un approccio alternativo potrebbe essere la creazione di un modello di struttura ad albero delle decisioni e quindi l'esame dei record erroneamente classificati come potenziali anomalie. Tuttavia, questo metodo presenterebbe maggiori difficoltà rispetto alla replica o all'automazione su larga scala).

Nodo Selezione funzioni

I problemi di data mining possono coinvolgere centinaia, se non migliaia, di campi che possono essere potenzialmente utilizzati come input. Di conseguenza, è possibile che l'analisi di quali campi o variabili includere in un modello richieda molto tempo e molti sforzi. Per circoscrivere le scelte, è possibile utilizzare l'algoritmo Selezione funzioni, che consente di identificare i campi più importanti per una determinata analisi. Per esempio, se si sta tentando di prevedere risultati clinici in base a una serie di fattori, quali fattori è più probabile che siano importanti?

La selezione delle funzioni include tre passaggi:

- **Screening.** Rimuove input e record o casi non importanti o problematici, quali campi di input con troppi valori mancanti o che presentano una variazione troppo grande o troppo piccola per risultare utili.
- **Classificazione.** Ordina gli input restanti e li classifica in base all'importanza.
- **Selezione.** Identifica il sottoinsieme di funzioni da utilizzare nei modelli successivi — ad esempio, conservando solo gli input più importanti e filtrando o escludendo tutti gli altri.

In un'epoca in cui molte organizzazioni sono sovraccaricate di dati, i vantaggi offerti della selezione delle funzioni per la semplificazione e l'accelerazione del processo di modellazione possono essere sostanziali.

Concentrando l'attenzione rapidamente sui campi più importanti, è possibile ridurre il numero di calcoli necessari, individuare più facilmente relazioni piccole ma importanti che potrebbero altrimenti passare inosservate e, in ultima analisi, ottenere modelli più semplici, più accurati e più facilmente spiegabili. Riducendo il numero di campi utilizzati nel modello, si scoprirà che è possibile ridurre il numero di dati raccolti nelle iterazioni future, nonché abbreviare i tempi di calcolo del punteggio.

Esempio. Un gestore telefonico dispone di un data warehouse contenente informazioni sulle risposte a una speciale promozione da parte di 5.000 clienti della società. I dati comprendono numerosi campi contenenti l'età, la professione, il reddito e le statistiche d'uso del telefono dei clienti. Tre campi obiettivo mostrano se il cliente ha aderito a ciascuna delle tre offerte che gli sono state proposte. La società desidera utilizzare tali dati per prevedere quali clienti hanno più probabilità di aderire ad offerte simili in futuro.

Requisiti. Un campo obiettivo (con il ruolo impostato su *Obiettivo*) singolo e più campi di input che si desidera sottoporre a screening o classificare rispetto all'obiettivo. Sia il campo obiettivo che i campi di input possono avere un livello di misurazione *Continuo* (intervallo numerico) o *Categoriale*.

Impostazioni del Modello di selezione funzioni

Le impostazioni nella scheda Modello includono le opzioni del modello standard e le impostazioni che consentono di ottimizzare i criteri per lo screening dei campi di input.

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.

Screening dei campi di input

Lo screening comporta la rimozione degli input o dei casi che non aggiungono alcuna informazione utile rispetto alla relazione input/obiettivo. Le opzioni di screening si basano sugli attributi del campo in questione senza considerare il potere predittivo relativo al campo obiettivo selezionato. I campi sottoposti a screening sono esclusi dai calcoli utilizzati per classificare gli input e, se lo si desidera, possono essere filtrati o rimossi dai dati utilizzati nella modellazione.

I campi possono essere sottoposti a screening in base ai seguenti criteri:

- **Percentuale massima dei valori mancanti.** Esegue lo screening dei campi con un numero troppo elevato di valori mancanti, espresso come percentuale del numero totale di record. I campi con una percentuale elevata di valori mancanti forniscono scarse informazioni predittive.
- **Percentuale massima di record in una categoria singola.** Esegue lo screening dei campi che hanno troppi record che rientrano nella stessa categoria rispetto al numero totale di record. Per esempio, se il 95% dei clienti nel database guida lo stesso tipo di auto, l'inclusione di questa informazione non è utile per distinguere un cliente da quello successivo. Tutti i campi che superano il numero massimo vengono sottoposti a screening. Questa opzione si applica solo ai campi categoriali.
- **Numero massimo di categorie come percentuale di record.** Esegue lo screening dei campi con troppe categorie rispetto al numero totale di record. Se un'elevata percentuale delle categorie contiene solo un unico caso, l'utilizzo del campo potrebbe essere limitato. Per esempio, se ogni cliente indossa un cappello diverso, è poco probabile che questa informazione sia utile per la modellazione di schemi comportamentali. Questa opzione si applica solo ai campi categoriali.
- **Coefficiente di variazione minimo.** Esegue lo screening dei campi con un coefficiente di varianza minore o uguale al minimo specificato. Questa misura corrisponde al rapporto tra la deviazione standard del campo di input e la media del campo di input. Se questo valore è prossimo allo zero, i valori della variabile non presentano una grande variabilità. Questa opzione si applica solo ai campi continui (intervallo numerico).
- **Deviazione standard minima.** Esegue lo screening dei campi con deviazione standard minore o uguale al minimo specificato. Questa opzione si applica solo ai campi continui (intervallo numerico).

Record con dati mancanti. I record o i casi con valori mancanti per il campo obiettivo, oppure con valori mancanti per tutti gli input, vengono esclusi automaticamente da tutti i calcoli utilizzati nelle classificazioni.

Opzioni di selezione funzioni

La scheda Opzioni consente di specificare le impostazioni di default per la selezione o l'esclusione dei campi di input nel nugget del modello. In seguito è possibile aggiungere il modello a un flusso per selezionare un sottoinsieme di campi da utilizzare nelle successive operazioni di generazione dei modelli. In alternativa, è possibile ignorare queste impostazioni selezionando o deselezionando campi aggiuntivi nel browser dei modelli dopo aver generato il modello. Tuttavia, le impostazioni di default consentono di applicare il nugget del modello senza ulteriori modifiche, il che può rivelarsi particolarmente utile ai fini dello script.

Per ulteriori informazioni, consultare l'argomento “Risultati del Modello di selezione funzioni” a pagina 56.

Sono disponibili le seguenti opzioni:

Tutti i campi classificati. Seleziona i campi in base alla loro classificazione come *importante*, *marginale* o *non importante*. È possibile modificare l'etichetta di ogni classificazione nonché i valori di interruzione utilizzati per assegnare i record all'uno o all'altro rango.

Primi N campi. Seleziona i primi *N* campi in base all'importanza.

Importanza maggiore di. Seleziona tutti i campi con importanza maggiore del valore specificato.

Il campo obiettivo viene sempre mantenuto indipendentemente dalla selezione.

Opzioni di classificazione dell'importanza

Tutti i categoriali. Quando tutti gli input ed i target sono relativi alla categoria, è possibile classificare l'importanza in base ad una delle quattro misure seguenti:

- **Chi-quadrato di Pearson.** Verifica l'indipendenza dell'obiettivo e dell'input senza indicare l'intensità o la direzione di qualsiasi relazione esistente.
- **Chi-quadrato del rapporto di verosimiglianza.** Simile al chi-quadrato di Pearson, ma verifica anche l'indipendenza obiettivo-input.
- **V di Cramer** Una misura di associazione basata sulla statistica del chi-quadrato di Pearson. I valori variano da 0, che indica nessuna associazione, a 1, che indica perfetta associazione.
- **Lambda.** Una misura di associazione che riflette la riduzione proporzionale nell'errore quando la variabile viene utilizzata per prevedere il valore obiettivo. Un valore di 1 indica che il campo di input prevede perfettamente l'obiettivo, mentre un valore di 0 significa che l'input non fornisce informazioni utili sull'obiettivo.

Alcuni categoriali. Quando alcuni —ma non tutti— gli input sono relativi alla categoria ed anche il target è relativo alla categoria, è possibile classificare l'importanza in base al chi-quadrato di Pearson o del rapporto di verosimiglianza. Il coefficiente *V* di Cramer e lambda non sono disponibili a meno che tutti gli input non siano categoriali.

Categoriale e continuo. Quando si classifica un input categoriale rispetto a un target continuo o viceversa (l'uno o l'altro sono categoriali ma non entrambi), viene utilizzata la statistica *F*.

Entrambi continui. Quando si classifica un input continuo rispetto a un target continuo, viene utilizzata la statistica *t* basata sul coefficiente di correlazione.

Nugget del modello Selezione funzioni

I nugget del modello Selezione funzioni visualizzano l'importanza di ciascun input rispetto a un obiettivo selezionato, secondo la classificazione del nodo Selezione funzioni. Vengono elencati anche i campi sottoposti a screening prima della classificazione. Per ulteriori informazioni, consultare l'argomento "Nodo Selezione funzioni" a pagina 53.

Quando si esegue un flusso contenente un nugget del modello Selezione funzioni, il modello agisce da filtro per mantenere solo gli input selezionati, come indicato dalla selezione corrente nella scheda Modello. Per esempio, è possibile selezionare tutti i campi classificati come importanti (una delle opzioni di default) o selezionare manualmente un sottoinsieme di campi nella scheda Modello. Anche il campo obiettivo viene mantenuto indipendentemente dalla selezione. Tutti gli altri campi vengono esclusi.

Il filtro si basa solo sul nome campo; per esempio, se si seleziona *età* e *reddito*, verranno mantenuti tutti i campi corrispondenti a questi nomi. Il modello non aggiorna le classificazioni dei campi in base ai nuovi dati, ma si limita a filtrare i campi in base ai nomi selezionati. Per questo motivo, è necessario prestare attenzione nell'applicazione del modello a dati nuovi o aggiornati. In caso di dubbio, si consiglia di rigenerare il modello.

Risultati del Modello di selezione funzioni

La scheda modello di un nugget del modello Selezione funzioni visualizza la classifica e l'importanza di tutti gli input nel riquadro superiore e consente di selezionare i campi per il filtro utilizzando le caselle di controllo nella colonna a sinistra. Quando viene eseguito il flusso, vengono mantenuti solo i campi selezionati; gli altri campi vengono ignorati. Le selezioni di default sono basate sulle opzioni specificate nel nodo di creazione del modello, ma è possibile selezionare o deselezionare campi aggiuntivi, se necessario.

Nel riquadro in basso sono elencati gli input che sono stati esclusi dalla classificazione in base alla percentuale di valori mancanti o ad altri criteri specificati nel nodo Modelli. Come per i campi classificati, è possibile scegliere di includere o scartare questi campi utilizzando le caselle di controllo nella colonna di sinistra. Per ulteriori informazioni, consultare l'argomento "Impostazioni del Modello di selezione funzioni" a pagina 54.

- Per ordinare l'elenco per rango, nome campo, importanza o in base a qualsiasi altra colonna visualizzata, fare clic sull'intestazione di colonna. In alternativa, per utilizzare la barra degli strumenti, selezionare l'elemento desiderato dall'elenco Ordina per e utilizzare le frecce su e giù per modificare la direzione dell'ordinamento.
- È possibile utilizzare la barra degli strumenti per selezionare o annullare la selezione di tutti i campi e per accedere alla finestra di dialogo Seleziona campi, che consente di selezionare i campi in base alla classifica o all'importanza. È anche possibile premere i tasti Maiusc e Ctrl mentre si fa clic sui campi per estendere la selezione e utilizzare la barra spaziatrice per attivare o disattivare la visualizzazione di un gruppo di campi selezionati. Per ulteriori informazioni, consultare l'argomento "Selezione dei campi per importanza".
- I valori di soglia per la classificazione degli input come importante, marginale o non importante vengono visualizzati nella legenda sotto alla tabella. Questi valori sono specificati nel nodo Modelli. Per ulteriori informazioni, consultare l'argomento "Opzioni di selezione funzioni" a pagina 55.

Selezione dei campi per importanza

Quando viene determinato il punteggio dei dati utilizzando un nugget del modello Selezione funzioni, tutti i campi selezionati dall'elenco dei campi classificati o sottoposti a screening, in base all'indicazione delle caselle di controllo nella colonna di sinistra, verranno mantenuti. Gli altri campi verranno scartati. Per modificare la selezione, è possibile utilizzare la barra degli strumenti per accedere alla finestra di dialogo Seleziona campi, che consente di selezionare i campi in base alla classifica o all'importanza.

Tutti i campi contrassegnati. Seleziona tutti i campi contrassegnati come importante, marginale o non importante.

Primi N campi. Consente di selezionare i primi *N* campi in base all'importanza.

Importanza maggiore di. Seleziona tutti i campi con importanza maggiore della soglia specificata.

Generazione di un filtro da un Modello di selezione funzioni

In base ai risultati del modello Selezione funzioni, è possibile utilizzare la finestra di dialogo Genera filtro da funzione per generare uno o più nodi filtro che includono o escludono sottoinsiemi di campi in base all'importanza relativa rispetto all'obiettivo specificato. Sebbene il nugget del modello possa essere utilizzato anche come filtro, questa funzione offre la flessibilità di sperimentare con diversi sottoinsiemi di campi senza copiare o modificare il modello. Il campo obiettivo viene sempre mantenuto dal filtro, indipendentemente dalla selezione di inclusione o esclusione.

Includi/Escludi È possibile scegliere di includere o escludere campi — ad esempio, includere i primi 10 campi o escludere tutti i campi contrassegnati come non importanti.

Campi selezionati. Include o esclude tutti i campi attualmente selezionati nella tabella.

Tutti i campi contrassegnati. Seleziona tutti i campi contrassegnati come importante, marginale o non importante.

Primi N campi. Consente di selezionare i primi *N* campi in base all'importanza.

Importanza maggiore di. Seleziona tutti i campi con importanza maggiore della soglia specificata.

Nodo Rilevamento anomalie

I modelli di rilevamento delle anomalie vengono utilizzati per identificare valori anomali, o casi insoliti, nei dati. A differenza di altri metodi di modellazione che archiviano regole su casi insoliti, i modelli di rilevamento delle anomalie archiviano informazioni su ciò che si intende per comportamento normale. Questo consente di identificare eventuali valori anomali, anche se non sono conformi ad alcuno schema noto, e ciò può essere particolarmente utile nelle applicazioni, per esempio per il rilevamento di comportamenti fraudolenti, in cui possono emergere continuamente nuovi schemi. Il rilevamento delle anomalie è un metodo non supervisionato, ovvero non richiede un insieme di dati di addestramento contenente casi noti di comportamenti fraudolenti come punto di partenza.

Mentre i metodi tradizionali di identificazione dei valori anomali controllano generalmente una o due variabili alla volta, il rilevamento delle anomalie può esaminare moltissimi campi per identificare cluster o gruppi di peer contenenti record simili. È quindi possibile confrontare ogni record con gli altri nel rispettivo gruppo di peer per identificare possibili anomalie. Più un caso è lontano dal centro normale, maggiore è la probabilità che questo caso sia insolito. Per esempio, l'algoritmo potrebbe raggruppare i record in tre cluster distinti e contrassegnare quelli che cadono lontano dal centro di ogni cluster.

A ogni record viene assegnato un indice di anomalie, che corrisponde al rapporto tra l'indice di deviazione del gruppo rispetto alla sua media e il cluster a cui appartiene il caso. Maggiore è il valore di tale indice, maggiore è la deviazione del caso rispetto alla media. In circostanze normali, i casi con un indice di anomalie inferiore a 1 o persino a 1,5 non sono considerati anomalie perché la deviazione corrisponde pressoché alla media o è di poco superiore Tuttavia, i casi con un valore di indice superiore a 2 potrebbero essere delle anomalie, poiché la deviazione è almeno doppia rispetto alla media.

Il rilevamento delle anomalie è un metodo esploratorio progettato per il rilevamento rapido di casi o record insoliti che potrebbero essere buoni candidati per un'ulteriore analisi. Tali casi e record dovrebbero essere considerati come anomalie *sospette* che, a un esame più attento, potrebbero non rivelarsi tali. Si

potrebbe per esempio scoprire che un record è perfettamente valido, ma decidere di escluderlo dai dati per la creazione del modello. Diversamente, se l'algoritmo rivela ripetutamente false anomalie, la causa potrebbe essere un errore o un valore falsato nel processo di raccolta dei dati.

Si noti che il rilevamento delle anomalie individua i record o i casi insoliti attraverso l'analisi dei cluster basata sull'insieme di campi selezionati nel modello, senza considerare alcun campo obiettivo (dipendente) specifico e indipendentemente dal fatto che questi campi siano pertinenti allo schema che si sta tentando di prevedere. Per questo motivo, l'utente potrebbe voler utilizzare il rilevamento di anomalie in combinazione con la selezione delle funzioni o un'altra tecnica per sottoporre a screening e classificare i campi. Per esempio, è possibile utilizzare la selezione delle funzioni per individuare i campi più importanti relativi a un obiettivo specifico e quindi utilizzare il rilevamento di anomalie per individuare i record più insoliti rispetto a tali campi (un approccio alternativo potrebbe essere la creazione di un modello di struttura ad albero delle decisioni e quindi l'esame dei record erroneamente classificati come potenziali anomalie. Tuttavia, questo metodo presenterebbe maggiori difficoltà rispetto alla replica o all'automazione su larga scala).

Esempio. Nello screening delle concessioni per lo sviluppo agricolo alla ricerca di possibili casi di frode, il rilevamento delle anomalie può essere utilizzato per scoprire le deviazioni dalla norma, grazie all'individuazione dei record anomali e che richiedono ulteriori analisi. L'interesse viene innanzitutto focalizzato sulle richieste di concessioni con una richiesta di denaro che sembra essere troppo elevata (o troppo esigua) per il tipo e le dimensioni dell'azienda agricola.

Requisiti. Uno o più campi di input. Si noti che solo i campi il cui ruolo è impostato su **Input** con un nodo origine o Tipo possono essere utilizzati come input. I campi obiettivo (ruolo impostato su **Obiettivo** o **Entrambi**) vengono ignorati.

Efficacia. Contrassegnando le caselle che *non* sono conformi a un insieme di regole note anziché quelle che lo sono, i modelli Rilevamento anomalie consentono di individuare i casi insoliti anche quando non seguono schemi già noti. Se utilizzato in combinazione con Selezione funzioni, il Rilevamento anomalie consente di sottoporre a screening grandi quantità di dati per individuare i record di maggior interesse in modo relativamente rapido.

Opzioni del modello Rilevamento anomalie

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.

Determina il valore di interruzione dell'anomalia in base a. Specifica il metodo utilizzato per determinare il valore di interruzione per contrassegnare le anomalie. Sono disponibili le seguenti opzioni:

- **Livello di indice di anomalie minimo.** Specifica il valore di interruzione minimo per contrassegnare le anomalie. I record che raggiungono o superano questa soglia vengono contrassegnati.
- **Percentuale dei record più anomali nei dati di addestramento.** Imposta automaticamente la soglia a un livello che contrassegna la percentuale specificata di record nei dati di addestramento. L'interruzione risultante viene inclusa come parametro nel modello. Si noti che questa opzione determina come viene impostato il valore di interruzione, *non* la percentuale effettiva di record da contrassegnare durante il calcolo del punteggio. I risultati effettivi del calcolo del punteggio possono variare in base ai dati.
- **Numero di record più anomali nei dati di addestramento.** Imposta automaticamente la soglia a un livello che contrassegna il numero specificato di record nei dati di addestramento. La soglia risultante viene inclusa come parametro nel modello. Si noti che questa opzione determina come viene impostato il valore di interruzione, *non* il numero specifico di record da contrassegnare durante il calcolo del punteggio. I risultati effettivi del calcolo del punteggio possono variare in base ai dati.

Nota: indipendentemente dalla sua modalità di determinazione, il valore di interruzione non influisce sul valore dell'indice di anomalie sottostante riportato per ciascun record, ma specifica semplicemente la

soglia per contrassegnare i record come anomali durante la stima o il calcolo del punteggio del modello. Se in un secondo tempo si desidera esaminare un numero maggiore o minore di record, è possibile utilizzare un nodo Seleziona per individuare un sottoinsieme di record in base al valore dell'indice delle anomalie (`$O-AnomalyIndex > X`).

Numero di campi di anomalie da inserire nel report. Specifica il numero di campi da inserire nel report come indicazione del motivo per il quale un determinato record è contrassegnato come anomalia. Vengono inseriti nel report i campi più anomali, cioè quelli che mostrano la maggior deviazione rispetto alla norma del campo per il cluster al quale il record è assegnato.

Opzioni avanzate del rilevamento anomalie

Per specificare le opzioni dei valori mancanti e le altre impostazioni, nella scheda Livello avanzato impostare la modalità su **Livello avanzato**.

Coefficiente di correzione. Valore utilizzato per bilanciare il peso relativo attribuito ai campi continui (intervallo numerico) e categoriali nel calcolo della distanza. Valori più grandi aumentano l'influenza dei campi continui. Questo valore deve essere diverso da zero.

Calcola automaticamente il numero dei gruppi di peer. Il Rilevamento anomalie può essere utilizzato per analizzare rapidamente un gran numero di soluzioni possibili per scegliere il numero di gruppi di peer ottimale per i dati di addestramento. È possibile estendere o restringere l'intervallo impostando il numero minimo e massimo di gruppi di peer. Valori maggiori consentiranno al sistema di esplorare un intervallo più esteso di soluzioni possibili; tuttavia, sarà necessario un tempo di elaborazione maggiore.

Specifica il numero dei gruppi di peer. Se si conosce il numero di cluster da includere nel modello, selezionare questa opzione e specificare il numero di gruppi di peer. Selezionando questa opzione si ottengono generalmente prestazioni migliori.

Livello e rapporto di rumore. Queste impostazioni determinano come saranno trattati i valori anomali durante il raggruppamento tramite cluster in due fasi. Nella prima fase viene utilizzata una struttura ad albero CF (cluster feature) per condensare i dati di un enorme numero di singoli record in un numero gestibile di cluster. La struttura ad albero viene creata in base alle misure di similarità; i nodi della struttura ad albero che contengono troppi record vengono suddivisi in nodi figlio. Nella seconda fase inizia il raggruppamento tramite cluster gerarchico sui nodi terminali della struttura ad albero CF. La gestione del rumore viene attivata nel primo passaggio di dati e disattivata nel secondo passaggio. I casi nel cluster rumore del primo passaggio di dati vengono assegnati ai cluster regolari nel secondo passaggio di dati.

- **Livello di rumore.** Specificare un valore compreso tra 0 e 0.5. Questa impostazione è rilevante solo se la struttura ad albero CF viene riempita durante la fase di espansione, a indicare che non è possibile accettare altri casi in un nodo foglia né suddividere alcun nodo foglia.

  Se la struttura ad albero CF viene riempita e il livello di rumore è impostato su 0, la soglia verrà aumentata e la struttura ad albero CF ingrandita di nuovo con tutti i casi. Dopo il raggruppamento finale, i valori non assegnati a un cluster vengono definiti valori anomali. Al cluster dei valori anomali viene assegnato il numero di identificazione –1. Il cluster dei valori anomali non è incluso nel conteggio del numero di cluster; se vengono specificati *n* cluster e la gestione del rumore, l'algoritmo genera *n* cluster ed un cluster rumore. In termini pratici, aumentando questo valore si conferisce all'algoritmo una maggior ampiezza per inserire i record insoliti nella struttura ad albero anziché assegnarli a un cluster di valori anomali separato.

  Se la struttura ad albero CF viene riempita e il livello di rumore è maggiore di 0, la struttura ad albero CF verrà ingrandita di nuovo dopo l'inserimento dei dati delle foglie sparse nelle proprie foglie di rumore. Una foglia si considera sparsa se il rapporto tra il numero di casi nella foglia sparsa e il numero di casi della foglia più grande è inferiore al livello di rumore. Una volta ingrandita la struttura ad albero, i valori anomali verranno inseriti nella struttura ad albero CF, se possibile. In caso contrario, i valori anomali vengono scartati per la seconda fase di raggruppamento tramite cluster.

- **Rapporto di rumore.** Specifica la parte di memoria allocata per il componente da utilizzare per la memorizzazione del rumore. Questo valore è compreso tra 0,0 e 0,5. Se l'inserimento di un caso specifico in una foglia della struttura ad albero determina un livello di precisione inferiore alla soglia, la foglia non viene suddivisa. Se il livello di precisione supera la soglia, la foglia viene suddivisa, aggiungendo un altro piccolo cluster alla struttura ad albero CF. In termini pratici, aumentando questa impostazione l'algoritmo potrebbe gravitare più rapidamente verso una struttura ad albero più semplice.

Assegna i valori mancanti. Per i campi continui, sostituisce i valori mancanti con la media dei campi. Per i campi categoriali, le categorie mancanti vengono combinate e trattate come una categoria valida. Se questa opzione è deselezionata, i record con valori mancanti vengono esclusi dall'analisi.

Nugget del modello Rilevamento anomalie

I nugget del modello Rilevamento anomalie contengono tutte le informazioni intercettate dal modello Rilevamento anomalie, nonché le informazioni sui dati di addestramento e sull'elaborazione della stima.

Quando si esegue un flusso contenente un nugget del modello Rilevamento anomalie, al flusso vengono aggiunti alcuni nuovi campi, a seconda delle selezioni effettuate nella scheda Impostazioni del nugget del modello. Per ulteriori informazioni, consultare l'argomento "Impostazioni del modello Rilevamento anomalie" a pagina 61. I nomi dei nuovi campi sono basati sul nome del modello, preceduti da *$O*, come riepilogato nella seguente tabella.

Tabella 6. Generazione del nome del nuovo campo.

Nome del campo	Descrizione
$O-Anomalia	Campo flag che indica se il record è anomalo o meno.
$O-IndiceAnomalie	Valore dell'indice delle anomalie del record.
$O-GruppoPeer	Specifica il gruppo di peer al quale è assegnato il record.
$O-Campo-n	Nome del campo *n* più anomalo in termini di deviazione rispetto alla norma del cluster.
$O-ImpattoCampo-n	Indice di deviazione variabile per il campo. Questo valore misura la deviazione rispetto alla norma del campo per il cluster al quale è assegnato il record.

Se lo si desidera, è possibile eliminare i punteggi per i record non anomali, in modo da facilitare la lettura dei risultati. Per ulteriori informazioni, consultare l'argomento "Impostazioni del modello Rilevamento anomalie" a pagina 61.

Dettagli del modello Rilevamento anomalie

La scheda Modello di un un modello Rilevamento anomalie generato visualizza informazioni sui gruppi di peer del modello.

Si noti che le dimensioni dei gruppi di peer e le statistiche inserite nei report sono stime basate sui dati di addestramento e possono differire leggermente dai risultati di calcolo del punteggio effettivi, anche se eseguite sugli stessi dati.

Riepilogo del modello Rilevamento anomalie

La scheda Riepilogo di un nugget del modello Rilevamento anomalie visualizza informazioni sui campi, le impostazioni di creazione e l'elaborazione della stima del modello. Viene inoltre visualizzato il numero di gruppi di peer, nonché il valore di interruzione utilizzato per contrassegnare i record come anomali.

Impostazioni del modello Rilevamento anomalie

La scheda Impostazioni consente di specificare le opzioni per il calcolo del punteggio del nugget del modello.

Indica i record anomali con. Specifica come vengono trattati i record anomali nell'output.

- **Contrassegna con flag e indicizza.** Crea un campo flag impostato su *Vero* per tutti i record che superano il valore di interruzione incluso nel modello. In un campo separato viene anche indicato l'indice delle anomalie di ogni record. Per ulteriori informazioni, consultare l'argomento "Opzioni del modello Rilevamento anomalie" a pagina 58.
- **Solo flag.** Crea un campo flag ma senza segnalare l'indice delle anomalie di ogni record.
- **Solo indice** Segnala l'indice delle anomalie senza creare un campo flag.

Numero di campi di anomalie da inserire nel report. Specifica il numero di campi da inserire nel report come indicazione del motivo per il quale un determinato record è contrassegnato come anomalia. Vengono inseriti nel report i campi più anomali, cioè quelli che mostrano la maggior deviazione rispetto alla norma del campo per il cluster al quale il record è assegnato.

Scarta record. Selezionare questa opzione per scartare tutti i record non anomali dal flusso, per facilitare l'analisi delle anomalie potenziali in qualsiasi nodo downstream. In alternativa, è possibile scegliere di scartare tutti i record anomali per limitare le analisi successive ai record non contrassegnati come potenziali anomalie in base al modello.

Nota: a causa di lievi differenze di arrotondamento, il numero effettivo di record contrassegnati durante il calcolo del punteggio potrebbe non essere identico a quello dei record contrassegnati durante l'addestramento del modello, anche se l'operazione viene eseguita sugli stessi dati.

Capitolo 5. Nodi Modelli automatici

I nodi di modellazione automatici stimano e confrontano diversi metodi di modellazione, consentendo di sperimentare approcci differenti in una singola esecuzione di modellazione. È possibile selezionare gli algoritmi di modellazione da utilizzare e le opzioni specifiche per ciascuno di essi, ivi comprese combinazioni che altrimenti si escluderebbero a vicenda. Per esempio, invece di dover scegliere tra i metodi rapido, dinamico e prune per una rete neurale, è possibile provare a utilizzarli tutti. Il nodo analizza ogni possibile combinazione di opzioni, classifica ogni modello candidato in base alle misure specificate dall'utente e salva i migliori per utilizzarli nel calcolo del punteggio o per ulteriori analisi.

È possibile scegliere fra tre nodi Modelli automatici, a seconda delle esigenze dell'analisi da eseguire:

Il nodo Classificatore automatico crea e confronta svariati tipi di modelli per risultati binari (sì o no, abbandono oppure no e così via), consentendo di scegliere l'approccio migliore per una determinata analisi. Sono supportati numerosi algoritmi di modellazione ed è possibile selezionare i metodi da utilizzare, le opzioni specifiche per ognuno di essi e i criteri per confrontare i risultati. Il nodo genera un insieme di modelli basato sulle opzioni specificate e classifica i candidati migliori in base ai criteri indicati.

Il nodo Numerico automatico stima e confronta i modelli per i risultati di intervalli numerici continui utilizzando svariati metodi. Il nodo funziona in modo analogo al nodo Classificatore automatico e consente di scegliere gli algoritmi da utilizzare e di sperimentare più combinazioni di opzioni in un singolo passaggio di modellazione. Gli algoritmi supportati includono reti neurali, C&R Tree, CHAID, regressione lineare, regressione lineare generalizzata e SVM (Support Vector Machine). I modelli si possono confrontare in base a correlazione, errore relativo o numero di variabili utilizzato.

Il nodo Cluster automatico stima e confronta i modelli di cluster che identificano gruppi di record con caratteristiche simili. Il nodo funziona in modo analogo ad altri nodi Modelli automatici e consente di sperimentare varie combinazioni di opzioni in un singolo passaggio di modellazione. I modelli si possono confrontare utilizzando misure di base con cui tentare di filtrare e classificare l'utilità dei modelli di cluster e fornire una misura in base all'importanza di determinati campi.

I modelli migliori vengono salvati in un singolo nugget del modello composito, per consentirne il confronto e per determinare i modelli da utilizzare per il calcolo del punteggio.

- Solo per gli obiettivi binari, nominali e numerici è possibile selezionare più modelli per il calcolo del punteggio e combinare i punteggi nei risultati di un insieme del modello. Se si combinano le previsioni di più modelli è possibile superare le limitazioni dei singoli modelli e ottenere così un grado di precisione complessivo migliore rispetto a quello ottenibile da ciascun modello.
- Se lo si desidera, è possibile decidere di eseguire il drill-down dei risultati e generare nodi Modelli o nugget del modello per ognuno dei modelli che si desidera utilizzare o sottoporre a un'analisi più approfondita.

Modelli e tempo di esecuzione

A seconda dell'insieme di dati e del numero di modelli, l'esecuzione dei nodi Modelli automatici può richiedere ore o tempi persino più lunghi. Durante la selezione delle opzioni, fare attenzione al numero dei modelli prodotti. Quando possibile, si consiglia di pianificare l'esecuzione dei modelli durante le ore notturne o nei fine settimana, vale a dire nei momenti in cui le risorse del sistema sono in genere meno utilizzate.

- Se necessario, è possibile utilizzare un nodo Partizione o un nodo Campione per ridurre il numero dei record inclusi nel passaggio di addestramento iniziale. Dopo che è stata ridotta la scelta a pochi modelli candidati, è possibile ripristinare l'insieme di dati completo.
- Per limitare il numero dei campi di input, utilizzare Selezione funzioni. Per ulteriori informazioni, consultare l'argomento "Nodo Selezione funzioni" a pagina 53. In alternativa, è possibile utilizzare le esecuzioni della modellazione iniziali per identificare i campi e le opzioni da analizzare in modo più approfondito. Per esempio, se apparentemente i modelli con le migliori prestazioni utilizzano tutti gli stessi tre campi, questa è una valida indicazione del fatto che detti campi devono essere conservati.
- È possibile limitare il tempo impiegato per stimare un determinato modello e specificare le misure di valutazione utilizzate per sottoporre i modelli a screening e per classificarli.

Impostazioni degli algoritmi dei nodi Modelli automatici

Per ogni tipo di modello è possibile utilizzare le impostazioni di default oppure scegliere le opzioni relative ai singoli tipi di modello. Le opzioni specifiche sono simili a quelle disponibili nei singoli nodi Modelli, tranne per il fatto che nella maggior parte dei casi è possibile scegliere il numero di impostazioni desiderato da applicare anziché scegliere l'una o l'altra impostazione. Per esempio, se si confrontano dei modelli Rete neurale, è possibile scegliere vari metodi di addestramento diversi e provare ogni metodo con e senza un seme random. Saranno utilizzate tutte le combinazioni possibili delle opzioni selezionate e risulterà estremamente semplice generare molti modelli diversi in un singolo passaggio. Occorre tuttavia prestare attenzione, perché la scelta contemporanea di più impostazioni può determinare una rapidissima moltiplicazione del numero dei modelli.

Per scegliere le opzioni per ogni tipo di modello

1. Nel nodo Modelli automatici, selezionare la scheda **Livello avanzato**.
2. Fare clic nella colonna **Parametri modello** per il tipo di modello.
3. Selezionare **Specifica** dal menu a discesa.
4. Nella finestra di dialogo **Impostazioni algoritmo**, selezionare le opzioni della colonna **Opzioni**.

Nota: nella scheda Livello avanzato della finestra di dialogo **Impostazioni algoritmo** sono disponibili ulteriori funzioni.

Regole di arresto dei nodi Modelli automatici

Le regole di arresto specificate per i nodi Modelli automatici sono relative all'esecuzione globale del nodo e non all'arresto dei singoli modelli creati dal nodo.

Limita tempo di esecuzione globale a. (Solo modelli Rete neurale, Medie K, Kohonen, TwoStep, SVM, KNN, Rete di Bayes e C&R Tree) Interrompe l'esecuzione dopo un numero di ore specificato. Tutti i modelli generati fino a quel punto vengono inclusi nel nugget del modello, ma non vengono prodotti altri modelli.

Interrompi non appena vengono prodotti modelli validi. Interrompe l'esecuzione quando un modello soddisfa tutti i criteri specificati nella scheda Scarta (per il nodo Classificatore automatico o Cluster automatico) o nella scheda Modello (per il nodo Numerico automatico). Per ulteriori informazioni, consultare l'argomento "Opzioni della scheda Scarta del nodo Classificatore automatico" a pagina 69. Per ulteriori informazioni, consultare l'argomento "Opzioni di scarto del nodo Cluster automatico" a pagina 76.

Nodo Classificatore automatico

Il nodo Classificatore automatico stima e confronta i modelli per gli obiettivi nominali (insieme) o binari (sì/no(, utilizzando una serie di metodi differenti, consentendo di sperimentare diversi approcci in una singola esecuzione di modellazione. È possibile selezionare gli algoritmi da utilizzare e sperimentare con varie combinazioni di opzioni. Per esempio, invece di dover scegliere tra i metodi rapido, dinamico e prune per una rete neurale, è possibile provare a utilizzarli tutti. Il nodo analizza ogni possibile combinazione di opzioni, classifica ciascun modello candidato in base alle misure specificate dall'utente e salva i migliori per utilizzarli nel calcolo del punteggio o per ulteriori analisi. Per ulteriori informazioni, consultare l'argomento Capitolo 5, “Nodi Modelli automatici”, a pagina 63.

Esempio. Una società di vendita al dettaglio dispone di dati cronologici che tengono traccia delle offerte fatte ai clienti specifici nell'ambito delle campagne precedenti. La società desidera raggiungere risultati più redditizi associando un'offerta appropriata a ciascun cliente.

Requisiti. Un campo obiettivo con livello di misurazione *Nominale* o *Flag* (con il ruolo impostato su **Obiettivo**) e almeno un campo di input (con il ruolo impostato su **Input**). Per un campo flag, si presuppone che il valore *Vero* definito per il campo obiettivo rappresenti un risultato quando si calcolano profitti, guadagno cumulativo e statistiche correlate. I campi di input possono avere un livello di misurazione *Continuo* o *Categoriale*, con la limitazione che alcuni input possono non essere appropriati per determinati tipi di modelli. Ad esempio, i campi ordinali utilizzati come input nei modelli C&R Tree, CHAID e QUEST devono disporre di archiviazione numerica (non stringa) e saranno ignorati da tali modelli se specificati diversamente. Analogamente, i campi di input continui possono essere discretizzati in alcuni casi. I requisiti sono uguali a quelli necessari per l'utilizzo dei singoli nodi Modelli; per esempio, un modello Rete di Bayes funziona allo stesso modo indipendentemente dal fatto che sia generato dal nodo Rete di Bayes o Classificatore automatico.

Campi frequenza e peso. La frequenza e il peso vengono utilizzati per conferire maggiore importanza ad alcuni record rispetto ad altri, per esempio perché l'utente sa che il dataset di creazione sottorappresenta una sezione della popolazione padre (Peso) o perché un record rappresenta diversi casi identici (Frequenza). Se specificato, un campo frequenza può essere utilizzato dai modelli C&R Tree, CHAID, QUEST, Elenco di decisioni e Rete di Bayes. Un campo peso può essere utilizzato dai modelli C&RT, CHAID e C5.0. Gli altri tipi di modelli ignoreranno questi campi e genereranno comunque i modelli. I campi frequenza e peso sono utilizzati solo per la creazione del modello e non vengono considerati per la valutazione o il calcolo del punteggio dei modelli. Per ulteriori informazioni, consultare l'argomento “Utilizzo dei campi frequenza e peso.” a pagina 33.

Tipi di modello supportati

I tipi di modelli supportati includono Rete neurale, C&R Tree, QUEST, CHAID, C5.0, Regressione logistica, Elenco di decisioni, Rete di Bayes, Discriminante, Vicino più prossimo e SVM. Per ulteriori informazioni, consultare l'argomento “Opzioni avanzate del nodo Classificatore automatico” a pagina 67.

Opzioni del modello di nodo Classificatore automatico

La scheda Modello del nodo Classificatore automatico consente di specificare il numero di modelli da creare, insieme ai criteri utilizzati per confrontare i modelli.

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.

Utilizza dati partizionati. Se è definito un campo partizione, questa opzione garantisce che per la creazione del modello vengano utilizzati dati solo dalla partizione di addestramento.

Crea modelli di suddivisione. Crea un modello separato per ogni valore possibile dei campi di input che sono stati specificati come campi di suddivisione. Per ulteriori informazioni, consultare l'argomento “Creazione di modelli di suddivisione” a pagina 28.

Classifica modelli per. Specifica i criteri utilizzati per confrontare e classificare i modelli. Le opzioni comprendono precisione complessiva, area sotto la curva ROC, profitto, guadagno cumulativo e numero di campi. Si noti che tutte queste misure saranno disponibili nel report di riepilogo, indipendentemente da quella selezionata in questa sede.

Nota: Per un obiettivo (insieme) nominale, la classificazione è limitata a **Accuratezza globale** o **Numero di campi**.

Quando si calcolano profitti, guadagno cumulativo e statistiche correlate, si presuppone che il valore *Vero* definito per il campo obiettivo rappresenti un risultato.

- **Precisione globale** La percentuale di record che viene prevista correttamente dal modello rispetto al numero totale di record.
- **Area sotto la curva ROC** La curva ROC fornisce un indice della performance di un modello. Più in alto si trova la curva al di sopra della linea di riferimento, più preciso è il test.
- **Profitto (cumulato)** La somma dei profitti dei percentili cumulati (ordinata in termini di confidenza per la previsione), calcolata in base a specifici criteri di costo, entrate e peso. Generalmente, il profitto inizia vicino allo 0 per il percentile più alto, aumenta in modo costante, quindi diminuisce. I profitti in un modello efficace mostreranno un picco ben definito, riportato assieme al percentile in cui si verifica. In un modello che non fornisce informazioni, la curva dei profitti sarà relativamente diritta e potrà seguire un andamento crescente, decrescente o piatto a seconda della struttura costo/entrate pertinente.
- **Guadagno cumulativo (cumulato)** Il rapporto tra i risultati nei quantili cumulati e il campione globale (dove i quantili vengono ordinati in termini di confidenza per la previsione). Per esempio, un valore guadagno cumulativo pari a 3 per il quantile più alto indica un tasso di risultati tre volte superiore a quello del campione globale. Il guadagno cumulativo di un modello efficace dovrebbe iniziare ben al di sopra di 1.0 per i quantili superiori e diminuire in modo marcato verso il livello di 1.0 per i quantili inferiori. In un modello che non fornisce informazioni, il guadagno cumulativo si aggira intorno a 1.0.
- **Numero di campi** Classifica i modelli in base al numero di campi di input utilizzati.

Classifica modelli mediante. Se viene utilizzata una partizione, è possibile specificare se i ranghi si basano sull'insieme di dati di addestramento o sull'insieme di test. Quando gli insiemi di dati sono molto estesi, l'utilizzo di una partizione per eseguire uno screening preliminare dei modelli può migliorare notevolmente la performance.

Numero di modelli da utilizzare. Specifica il numero massimo di modelli che verranno elencati nel nugget del modello prodotto dal nodo. I modelli con la classificazione più alta vengono elencati in base al criterio di classificazione specificato. Si noti che l'incremento di questo limite potrebbe determinare una diminuzione delle performance. Il valore massimo consentito è 100.

Calcola importanza predittori. Per i modelli che generano una misura appropriata dell'importanza è possibile visualizzare un grafico che indica l'importanza relativa di ogni predittore nella stima del modello. In genere è opportuno concentrare la modellazione sui predittori più importanti e valutare se eliminare o ignorare quelli meno importanti. Si noti che l'importanza dei predittori può aumentare il tempo necessario per calcolare alcuni modelli e non è consigliabile considerarla se si desidera semplicemente effettuare un confronto generale tra molti modelli diversi. Essa risulta più utile quando l'analisi è stata ristretta a pochi modelli che si desidera analizzare in modo più approfondito. Per ulteriori informazioni, consultare l'argomento “Importanza predittore” a pagina 43.

Criteri di profitto. *Nota*: solo per obiettivi flag. Il profitto equivale alle entrate relative a ciascun record meno il costo del record. I profitti di un quantile sono semplicemente la somma dei profitti di tutti i record nel quantile. Si presuppone che i profitti vengano applicati solo ai risultati, mentre i costi vengono applicati a tutti i record.

- **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 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 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 grafici ROC).

Criteri di guadagno cumulativo. *Nota*: solo per obiettivi flag. Specifica il percentile da utilizzare per i calcoli del guadagno cumulativo. Si noti che è possibile modificare questo valore anche quando si confrontano i risultati. Per ulteriori informazioni, consultare l'argomento “Nugget del modello automatici” a pagina 77.

Opzioni avanzate del nodo Classificatore automatico

La scheda Livello avanzato del nodo Classificatore automatico consente di applicare una partizione (se disponibile), selezionare gli algoritmi da utilizzare e specificare le regole di arresto.

Seleziona modelli. Per impostazione predefinita tutti i modelli sono selezionati per essere creati; tuttavia se si dispone di Analytic Server, è possibile scegliere di limitare i modelli a quelli che è possibile eseguire su Analytic Server e preimpostarli in modo che creino modelli di suddivisione o sono pronti per elaborare dataset di grandi dimensioni.

Modelli utilizzati. Utilizzare le caselle di controllo nella colonna di sinistra per selezionare i tipi di modelli (algoritmi) da includere nel confronto. Più tipi di modelli vengono selezionati, più modelli verranno creati e più lunghi saranno i tempi di elaborazione.

Tipo di modello. Elenca gli algoritmi disponibili (vedere di seguito).

Parametri del modello. Per ogni tipo di modello è possibile utilizzare le impostazioni di default oppure selezionare **Specifica** per scegliere le opzioni relative ai singoli tipi di modello. Le opzioni specifiche sono simili a quelle disponibili nei singoli nodi Modelli, con la differenza che è possibile selezionare più opzioni o combinazioni di opzioni. Per esempio, se si confrontano modelli Rete neurale, invece di scegliere uno dei sei metodi di addestramento, è possibile scegliere tutti i metodi in modo da addestrare sei modelli in un unico passaggio.

Numero di modelli. Elenca il numero di modelli prodotti per ciascun algoritmo in base alle impostazioni correnti. Quando si sceglie una combinazione di opzioni, il numero di modelli può aumentare rapidamente; pertanto, si consiglia di fare molta attenzione a questo numero, soprattutto se si utilizzano insiemi di dati molto estesi.

Limita tempo massimo impiegato per creare un modello singolo. (Solo modelli Medie K, Kohonen, TwoStep, SVM, KNN, Rete di Bayes ed Elenco di decisioni) Imposta un limite di tempo massimo per ogni modello. Per esempio, se un modello richiede un tempo di addestramento più lungo del previsto a causa di qualche interazione complessa, grazie a questa opzione è possibile evitare che l'operazione ritardi e blocchi l'esecuzione di tutta la modellazione.

Nota: se il target è un campo nominale (insieme), l'opzione Elenco di decisioni non è disponibile.

Algoritmi supportati

Il nodo Rete neurale utilizza un modello semplificato del modo in cui il cervello umano elabora le informazioni. Funziona simulando un elevato numero di semplici unità di elaborazione interconnesse che assomigliano a versioni astratte di neuroni. Le reti neurali sono potenti strumenti di valutazione delle funzioni generali e richiedono una conoscenza statistica o matematica minima per l'addestramento o l'applicazione.

Il nodo C5.0 crea una struttura ad albero delle decisioni o un insieme di regole. Il modello suddivide il campione in base al campo che fornisce il massimo guadagno di informazioni a ogni livello. Il campo obiettivo deve essere categoriale. Sono consentite suddivisioni multiple in più di due sottogruppi.

Il nodo Struttura ad albero di classificazione e regressione (C&R) genera una struttura ad albero delle decisioni che consente di prevedere o classificare osservazioni future. Il metodo utilizza partizionamento ricorsivo per suddividere i record di addestramento in segmenti, riducendo l'impurità ad ogni passaggio. Un nodo della struttura ad albero è considerato "puro" quando il 100% dei casi nel nodo fa parte di una categoria specifica del campo obiettivo. I campi obiettivo e di input possono essere intervalli numerici o categoriali (nominali, ordinali o flag); tutte le suddivisioni sono binarie (solo due sottogruppi).

Il nodo QUEST offre un metodo di classificazione binario per la creazione di strutture ad albero delle decisioni, progettato per ridurre i tempi di elaborazione necessari per le analisi C&R Tree più complesse, riducendo inoltre la tendenza dei metodi per le strutture ad albero di classificazione a favorire gli input che consentono un numero maggiore di suddivisioni. I campi di input possono essere intervalli numerici (continui), ma il campo obiettivo deve essere categoriale. Tutte le suddivisioni sono binarie.

Il nodo CHAID genera una struttura ad albero delle decisioni utilizzando statistiche chi-quadrato per identificare suddivisioni ottimali. A differenza dei nodi C&R Tree e QUEST, il nodo CHAID può generare strutture ad albero non binarie e pertanto alcune suddivisioni possono avere più di due rami. I campi obiettivo e di input possono essere intervallo numerico (continui) o categoriali. Un CHAID completo è una modificazione di CHAID che esegue operazioni avanzate per l'analisi di tutte le suddivisioni possibili, ma richiede tempi di elaborazione maggiori.

La regressione logistica, una tecnica statistica che consente di classificare i record in base ai valori dei campi di input, è analoga alla regressione lineare ma, al posto di un intervallo numerico, prende un campo obiettivo categoriale.

Il nodo Elenco di decisioni identifica i sottogruppi o i segmenti che mostrano una probabilità maggiore o minore che si verifichi un determinato risultato binario rispetto alla popolazione globale. Per esempio, è possibile che si cerchino i clienti non a rischio di abbandono o quelli che più probabilmente rispondano in modo favorevole a una campagna. È possibile incorporare le proprie conoscenze di business nel modello aggiungendo propri segmenti personalizzati e visualizzando in anteprima modelli alternativi uno accanto all'altro per confrontarne i risultati. I modelli Elenco di decisioni consistono in un elenco di regole in cui ogni regola ha una condizione e un risultato. Le regole vengono applicate in ordine e la prima regola corrispondente determina il risultato.

Il nodo Rete bayesiana consente di generare un modello di probabilità combinando elementi osservati e registrati con conoscenze del mondo reale per stabilire la probabilità di occorrenze. Il nodo si concentra sulle reti TAN (Tree Augmented Naïve Bayes) e coperta di Markov, che sono prevalentemente utilizzate a scopo di classificazione.

L'analisi discriminante prevede presupposti più rigidi rispetto alla regressione logistica, ma può essere una valida alternativa o un complemento dell'analisi di regressione logistica quando vengono soddisfatti tali presupposti.

Il nodo Elemento vicino più prossimo *K* (KNN) associa un nuovo caso alla categoria o valore degli oggetti *K* più vicini ad esso nello spazio predittore, dove *K* è un numero intero. I casi simili sono vicini gli uni agli altri, mentre i casi dissimili sono distanti gli uni dagli altri.

Il nodo SVM (Support Vector Machine) consente di classificare i dati in uno di due gruppi senza sovradattamento. Il nodo SVM è particolarmente indicato per l'utilizzo con insiemi di dati di grandi dimensioni, cioè quelli con un elevato numero di campi di input.

Costi classificazione errata

In alcuni contesti, certi tipi di errori rappresentano un costo maggiore rispetto ad altri. Potrebbe essere più costoso, per esempio, classificare come a basso rischio chi richiede un fido ad alto rischio (un tipo di errore) di quanto non lo sia classificare come ad alto rischio chi chiede un fido a basso rischio (un tipo di errore diverso). I costi di errata classificazione permettono di specificare l'importanza relativa dei diversi tipi di errori di previsione.

Essenzialmente i costi di errata classificazione sono pesi applicati a risultati specifici. Tali pesi vengono fattorizzati nel modello e possono realmente modificare la previsione (come modo per proteggersi da errori che gravano sui costi).

Ad eccezione dei modelli C5.0, i costi di errata classificazione non sono applicati quando si calcola il punteggio di un modello e non vengono presi in considerazione quando si classificano o confrontano modelli con il nodo Classificatore automatico, un nodo Analisi o un grafico di valutazione. È possibile che un modello che include costi non generi un numero inferiore di errori rispetto a uno che non li include e che non si classifichi meglio in termini di precisione complessiva, ma è probabile che fornisca prestazioni migliori in termini pratici poiché include una distorsione incorporata a favore degli errori *meno costosi*.

La matrice dei costi mostra il costo per ciascuna possibile combinazione di categoria prevista e categoria effettiva. Per default, tutti i costi di errata classificazione sono impostati su 1.0. Per immettere valori di costi personalizzati, selezionare **Utilizza costi di errata classificazione** e immettere i valori personalizzati nella matrice dei costi.

Per cambiare un costo di errata classificazione, selezionare la cella corrispondente alla combinazione desiderata di valori previsti e valori effettivi, eliminare il contenuto esistente della cella e immettere il costo desiderato. I costi non sono simmetrici automaticamente. Se si imposta, per esempio, il costo dell'errata classificazione di *A* come *B* su 2.0 e non viene esplicitamente cambiato il costo dell'errata classificazione di *B* come *A*, esso manterrà il valore di default 1.0.

Opzioni della scheda Scarta del nodo Classificatore automatico

La scheda Scarta del nodo Classificatore automatico consente di scartare automaticamente i modelli che non soddisfano determinati criteri. Questi modelli non verranno inclusi nell'elenco del report di riepilogo.

È possibile specificare una soglia minima per la precisione globale, nonché una soglia massima per il numero delle variabili utilizzate nel modello. Inoltre, nel caso di obiettivi flag, è possibile specificare una soglia minima per guadagno cumulativo, profitto e area sotto la curva; guadagno cumulativo e profitto

vengono determinati come specificato nella scheda Modello. Per ulteriori informazioni, consultare l'argomento “Opzioni del modello di nodo Classificatore automatico” a pagina 65.

Se lo si desidera, è possibile configurare il nodo in modo che l'esecuzione venga interrotta alla prima generazione di un modello che soddisfa tutti i criteri specificati. Per ulteriori informazioni, consultare l'argomento “Regole di arresto dei nodi Modelli automatici” a pagina 64.

Opzioni di impostazione del nodo Classificatore automatico

La scheda Impostazioni del nodo Classificatore automatico consente di preconfigurare le opzioni tempo-punteggio disponibili nel nugget.

Metodo dell'insieme. Per gli obiettivi è possibile selezionare dai seguenti metodi dell'insieme:

- Voting
- Confronto ponderato con confidenza
- Confronto ponderato con propensione grezza (solo obiettivi flag)
- Confidenza più elevata vince
- Propensione grezza media (solo obiettivi flag)

Se il confronto risulta a pari merito, selezionare il valore utilizzando: 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.** (Solo obiettivi flag) Il valore pari merito previsto con la massima propensione assoluta, dove la propensione assoluta è calcolata come:

```
abs(0.5 - propensione) * 2
```

Nodo Numerico automatico

Il nodo Numerico automatico stima e confronta i modelli per i risultati di intervalli numerici continui utilizzando una serie di metodi differenti, consentendo di sperimentare approcci diversi in una singola esecuzione di modellazione. È possibile selezionare gli algoritmi da utilizzare e sperimentare con varie combinazioni di opzioni. Ad esempio, è possibile prevedere i valori delle abitazioni utilizzando i modelli rete neurale, regressione lineare, C&RT e CHAID per verificare quale di essi offre le prestazioni migliori ed è possibile provare diverse combinazioni dei metodi di regressione stepwise, in avanti ed all'indietro. Il nodo analizza ogni possibile combinazione di opzioni, classifica ogni modello candidato in base alle misure specificate dall'utente e salva i migliori per utilizzarli nel calcolo del punteggio o per ulteriori analisi. Per ulteriori informazioni, consultare l'argomento Capitolo 5, “Nodi Modelli automatici”, a pagina 63.

Esempio. Un'amministrazione comunale desidera effettuare una stima più precisa delle imposte sugli immobili e ritoccare, se necessario, i valori relativi a determinate proprietà senza doverle ispezionare tutte. Utilizzando il nodo Numerico automatico, l'analista può generare e confrontare numerosi modelli che prevedono i valori della proprietà in base al tipo di edificio, al quartiere, alle dimensioni e ad altri fattori noti.

Requisiti. Un solo campo obiettivo (con il ruolo impostato su **Obiettivo**) e almeno un campo di input (con il ruolo impostato su **Input**). L'obiettivo deve essere un campo continuo (intervallo numerico) quale *età* o *reddito*. I campi di input possono essere continui o categoriali, con la limitazione che alcuni input possono non essere appropriati per determinati tipi di modelli. Ad esempio, i modelli C&R Tree possono utilizzare campi stringa relativi alla categoria come input, mentre i modelli di regressione lineare non possono utilizzare tali campi che, se specificati, vengono ignorati. I requisiti sono analoghi a quelli

richiesti per l'utilizzo dei singoli nodi Modelli. Per esempio, un modello CHAID funziona allo stesso modo sia quando è generato dal nodo CHAID, sia quando è generato dal nodo Numerico automatico.

Campi frequenza e peso. La frequenza e il peso vengono utilizzati per conferire maggiore importanza ad alcuni record rispetto ad altri, per esempio perché l'utente sa che l'insieme di dati di creazione sottorappresenta una sezione della popolazione padre (Peso) o perché un record rappresenta diversi casi identici (Frequenza). Se specificato, un campo frequenza può essere utilizzato dagli algoritmi C&R Tree e CHAID. Un campo peso può essere utilizzato dagli algoritmi C&RT, CHAID, Regressione e GenLin. Gli altri tipi di modelli ignoreranno questi campi e genereranno comunque i modelli. I campi frequenza e peso sono utilizzati solo per la creazione del modello e non vengono considerati per la valutazione o il calcolo del punteggio dei modelli. Per ulteriori informazioni, consultare l'argomento “Utilizzo dei campi frequenza e peso.” a pagina 33.

Tipi di modello supportati

I tipi di modello supportati includono Rete neurale, C&R Tree, CHAID, Regressione, GenLin, Vicino più prossimo e SVM. Per ulteriori informazioni, consultare l'argomento “Opzioni avanzate del nodo Numerico automatico” a pagina 72.

Opzioni del modello di nodo Numerico automatico

La scheda Modello del nodo Numerico automatico consente di specificare il numero di modelli da salvare ed i criteri utilizzati per confrontare i modelli.

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.

Utilizza dati partizionati. Se è definito un campo partizione, questa opzione garantisce che per la creazione del modello vengano utilizzati dati solo dalla partizione di addestramento.

Crea modelli di suddivisione. Crea un modello separato per ogni valore possibile dei campi di input che sono stati specificati come campi di suddivisione. Per ulteriori informazioni, consultare l'argomento “Creazione di modelli di suddivisione” a pagina 28.

Classifica modelli per. Specifica i criteri utilizzati per confrontare i modelli.

- **Correlazione.** La correlazione di Pearson tra il valore osservato per ogni record e il valore previsto dal modello. La correlazione è una misura dell'associazione lineare tra due variabili, in cui i valori più vicini a 1 indicano una relazione più forte. I valori di correlazione sono compresi tra –1, che indica una relazione negativa perfetta, e +1 per una relazione positiva perfetta. Il valore 0 indica l'assenza di relazione lineare, mentre un modello con correlazione negativa avrebbe la relazione più debole in assoluto.
- **Numero di campi.** Il numero dei campi utilizzati come predittori nel modello. La scelta di modelli che utilizzano un numero inferiore di campi può semplificare la preparazione dei dati e, in alcuni casi, migliorare le prestazioni.
- **Errore relativo.** L'errore relativo è il rapporto tra la varianza dei valori osservati rispetto a quelli previsti dal modello e la varianza dei valori osservati rispetto alla media. In pratica, esso confronta l'efficacia delle prestazioni del modello rispetto a un modello **null** o **intercettazione** che restituisce semplicemente il valore medio del campo obiettivo come previsione. In un modello valido, questo valore deve essere inferiore a 1, a indicare che il modello è più preciso del modello null. Un modello con un errore relativo superiore a 1 è meno preciso del modello null e pertanto non è utile. Per i modelli di regressione lineare, l'errore relativo è uguale al quadrato della correlazione e non fornisce alcuna informazione nuova. Per i modelli non lineari, l'errore relativo è indipendente dalla correlazione e fornisce una misura aggiuntiva per la valutazione delle prestazioni del modello.

Classifica modelli mediante. Se viene utilizzata una partizione, è possibile specificare se i ranghi si basano sulla partizione di addestramento o sulla partizione di test. Quando gli insiemi di dati sono molto estesi, l'utilizzo di una partizione per eseguire uno screening preliminare dei modelli può migliorare notevolmente la performance.

Numero di modelli da utilizzare. Specifica il numero massimo di modelli che verranno visualizzati nel nugget del modello prodotto dal nodo. I modelli con la classificazione più alta vengono elencati in base al criterio di classificazione specificato. Aumentando questo limite, è possibile confrontare i risultati per ulteriori modelli, ma ciò potrebbe rallentare le prestazioni. Il valore massimo consentito è 100.

Calcola importanza predittori. Per i modelli che generano una misura appropriata dell'importanza è possibile visualizzare un grafico che indica l'importanza relativa di ogni predittore nella stima del modello. In genere è opportuno concentrare la modellazione sui predittori più importanti e valutare se eliminare o ignorare quelli meno importanti. Si noti che l'importanza dei predittori può aumentare il tempo necessario per calcolare alcuni modelli e non è consigliabile considerarla se si desidera semplicemente effettuare un confronto generale tra molti modelli diversi. Essa risulta più utile quando l'analisi è stata ristretta a pochi modelli che si desidera analizzare in modo più approfondito. Per ulteriori informazioni, consultare l'argomento “Importanza predittore” a pagina 43.

Non mantenere modelli se. Specifica i valori di soglia per la correlazione, l'errore relativo e il numero di campi utilizzati. I modelli che non soddisfano uno qualsiasi di questi criteri vengono scartati e non compariranno nel report di riepilogo.

- **La correlazione è inferiore a.** La correlazione minima (in termini di valore assoluto) di un modello da includere nel report di riepilogo.
- **Il numero di campi è maggiore di.** Il numero massimo di campi da utilizzare in qualsiasi modello da includere.
- **L'errore relativo è maggiore di.** L'errore relativo massimo per ogni modello da includere.

Se lo si desidera, è possibile configurare il nodo in modo che l'esecuzione venga interrotta alla prima generazione di un modello che soddisfa tutti i criteri specificati. Per ulteriori informazioni, consultare l'argomento “Regole di arresto dei nodi Modelli automatici” a pagina 64.

Opzioni avanzate del nodo Numerico automatico

La scheda Livello avanzato del nodo Numerico automatico consente di selezionare gli algoritmi e le opzioni da utilizzare e di specificare le regole di arresto.

Seleziona modelli. Per impostazione predefinita tutti i modelli sono selezionati per essere creati; tuttavia se si dispone di Analytic Server, è possibile scegliere di limitare i modelli a quelli che è possibile eseguire su Analytic Server e preimpostarli in modo che creino modelli di suddivisione o sono pronti per elaborare dataset di grandi dimensioni.

Modelli utilizzati. Utilizzare le caselle di controllo nella colonna di sinistra per selezionare i tipi di modelli (algoritmi) da includere nel confronto. Più tipi di modelli vengono selezionati, più modelli verranno creati e più lunghi saranno i tempi di elaborazione.

Tipo di modello. Elenca gli algoritmi disponibili (vedere di seguito).

Parametri del modello. Per ogni tipo di modello è possibile utilizzare le impostazioni di default oppure selezionare **Specifica** per scegliere le opzioni relative ai singoli tipi di modello. Le opzioni specifiche sono simili a quelle disponibili nei singoli nodi Modelli, con la differenza che è possibile selezionare più opzioni o combinazioni di opzioni. Per esempio, se si confrontano modelli Rete neurale, invece di scegliere uno dei sei metodi di addestramento, è possibile scegliere tutti i metodi in modo da addestrare sei modelli in un unico passaggio.

Numero di modelli. Elenca il numero di modelli prodotti per ciascun algoritmo in base alle impostazioni correnti. Quando si sceglie una combinazione di opzioni, il numero di modelli può aumentare rapidamente; pertanto, si consiglia di fare molta attenzione a questo numero, soprattutto se si utilizzano insiemi di dati molto estesi.

Limita tempo massimo impiegato per creare un modello singolo. (Solo modelli Medie K, Kohonen, TwoStep, SVM, KNN, Rete di Bayes ed Elenco di decisioni) Imposta un limite di tempo massimo per ogni modello. Per esempio, se un modello richiede un tempo di addestramento più lungo del previsto a causa di qualche interazione complessa, grazie a questa opzione è possibile evitare che l'operazione ritardi e blocchi l'esecuzione di tutta la modellazione.

Algoritmi supportati

Il nodo Rete neurale utilizza un modello semplificato del modo in cui il cervello umano elabora le informazioni. Funziona simulando un elevato numero di semplici unità di elaborazione interconnesse che assomigliano a versioni astratte di neuroni. Le reti neurali sono potenti strumenti di valutazione delle funzioni generali e richiedono una conoscenza statistica o matematica minima per l'addestramento o l'applicazione.

Il nodo Struttura ad albero di classificazione e regressione (C&R) genera una struttura ad albero delle decisioni che consente di prevedere o classificare osservazioni future. Il metodo utilizza partizionamento ricorsivo per suddividere i record di addestramento in segmenti, riducendo l'impurità ad ogni passaggio. Un nodo della struttura ad albero è considerato "puro" quando il 100% dei casi nel nodo fa parte di una categoria specifica del campo obiettivo. I campi obiettivo e di input possono essere intervalli numerici o categoriali (nominali, ordinali o flag); tutte le suddivisioni sono binarie (solo due sottogruppi).

Il nodo CHAID genera una struttura ad albero delle decisioni utilizzando statistiche chi-quadrato per identificare suddivisioni ottimali. A differenza dei nodi C&R Tree e QUEST, il nodo CHAID può generare strutture ad albero non binarie e pertanto alcune suddivisioni possono avere più di due rami. I campi obiettivo e di input possono essere intervallo numerico (continui) o categoriali. Un CHAID completo è una modificazione di CHAID che esegue operazioni avanzate per l'analisi di tutte le suddivisioni possibili, ma richiede tempi di elaborazione maggiori.

La regressione lineare è una tecnica statistica molto comune per riassumere i dati ed eseguire previsioni individuando un'area o una linea retta in grado di ridurre le discrepanze tra i valori di output previsti e quelli osservati.

Il modello Lineare generalizzato amplia il modello lineare generale in modo che la variabile dipendente venga linearmente correlata ai fattori e alle covariate tramite una funzione di collegamento specifica. Inoltre, il modello consente alla variabile dipendente di avere una distribuzione non normale. Copre la funzionalità di un grande numero di modelli statistici, inclusi modelli di regressione lineare, modelli di regressione logistica, modelli loglineari per dati dei conteggi e modelli di sopravvivenza censurati per intervallo.

Il nodo Elemento vicino più prossimo *K* (KNN) associa un nuovo caso alla categoria o valore degli oggetti *K* più vicini ad esso nello spazio predittore, dove *K* è un numero intero. I casi simili sono vicini gli uni agli altri, mentre i casi dissimili sono distanti gli uni dagli altri.

Il nodo SVM (Support Vector Machine) consente di classificare i dati in uno di due gruppi senza sovradattamento. Il nodo SVM è particolarmente indicato per l'utilizzo con insiemi di dati di grandi dimensioni, cioè quelli con un elevato numero di campi di input.

I modelli di regressione lineare prevedono un target continuo basato sulle relazioni lineari tra l'obiettivo e uno o più predittori.

Opzioni di impostazione del nodo Numerico automatico

La scheda Impostazioni del nodo Numerico automatico consente di preconfigurare le opzioni tempo-punteggio disponibili nel nugget.

Calcola errore standard. In caso di un target continuo (intervallo numerico), viene eseguito per default il calcolo dell'errore standard per calcolare la differenza fra i valori misurati o stimati e i valori veri e per evidenziare il grado di corrispondenza di tali stime.

Nodo Cluster automatico

Il nodo Cluster automatico stima e confronta i modelli di cluster che identificano gruppi di record con caratteristiche simili. Il nodo funziona in modo analogo ad altri nodi di modellazione automatici, consentendo di sperimentare più combinazioni di opzioni in un singolo passaggio di modellazione. I modelli si possono confrontare utilizzando misure di base con cui tentare di filtrare e classificare l'utilità dei modelli di cluster e fornire una misura in base all'importanza di determinati campi.

I modelli di cluster vengono spesso utilizzati per identificare gruppi che è possibile impiegare come input in analisi successive. Per esempio, si potrebbe desiderare di rivolgersi a gruppi di clienti in base a caratteristiche demografiche quali il reddito, o in base ai servizi acquistati in passato. Questo è possibile anche senza conoscere a priori i gruppi e le loro caratteristiche: è possibile che non si conosca il numero dei gruppi da cercare, o le caratteristiche da utilizzare nella loro definizione. I modelli di cluster vengono spesso definiti modelli di apprendimento non supervisionato, poiché non utilizzano un campo obiettivo e non restituiscono una previsione specifica che possa essere valutata come vera o falsa. Il valore di un modello di cluster è determinato dalla sua capacità di acquisire gruppi significativi all'interno dei dati e di fornire descrizioni utili di tali raggruppamenti. Per ulteriori informazioni, consultare Capitolo 11, “Modelli di cluster”, a pagina 213.

Requisiti. Uno o più campi che definiscano caratteristiche interessanti. I modelli di cluster non utilizzano i campi obiettivo nello stesso modo degli altri modelli, perché non effettuano previsioni specifiche che è possibile valutare come vere o false. Essi sono invece utilizzati per identificare gruppi di casi che potrebbero essere correlati. Per esempio, non è possibile utilizzare un modello di cluster per prevedere se un determinato cliente abbandonerà o risponderà a un'offerta. Con un modello di cluster si può tuttavia assegnare i clienti a dei gruppi in base alla loro tendenza a comportarsi in un modo o nell'altro. I campi peso e frequenza non sono utilizzati.

Campi di valutazione. Benché non vengano utilizzati obiettivi, è possibile specificare uno o più campi di valutazione da utilizzare nel confronto dei modelli. L'utilità di un modello di cluster si può valutare misurando la capacità dei cluster di operare una distinzione tra questi campi.

Tipi di modello supportati

I tipi di modelli supportati includono i modelli TwoStep, Medie K e Kohonen.

Opzioni del modello di nodo Cluster automatico

La scheda Modello del nodo Cluster automatico consente di specificare il numero di modelli da salvare, insieme ai criteri utilizzati per il confronto dei modelli.

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.

Utilizza dati partizionati. Se è definito un campo partizione, questa opzione garantisce che per la creazione del modello vengano utilizzati dati solo dalla partizione di addestramento.

Classifica modelli per. Specifica i criteri utilizzati per confrontare e classificare i modelli.

- **Silhouette.** Indice che consente la misurazione della coesione e della separazione dei cluster. Per ulteriori informazioni, vedere *Misura di classificazione silhouette* di seguito.
- **Numero di cluster.** Il numero dei cluster nel modello.
- **Dimensioni del cluster più piccolo.** La dimensione più ridotta di cluster.
- **Dimensioni del cluster più grande.** La dimensione più grande di cluster.
- **Cluster più piccolo/più grande.** Il rapporto tra le dimensioni del cluster più piccolo e le dimensioni di quello più grande.
- **Importanza.** L'importanza del campo **Valutazione** nella scheda **Campi**. Tenere presente che questo valore può essere calcolato solo se è stato specificato un campo **Valutazione**.

Classifica modelli mediante. Se viene utilizzata una partizione, è possibile specificare se i ranghi si basano sull'insieme di dati di addestramento o sull'insieme di test. Quando gli insiemi di dati sono molto estesi, l'utilizzo di una partizione per eseguire uno screening preliminare dei modelli può migliorare notevolmente la performance.

Numero di modelli da conservare. Specifica il numero massimo di modelli che verranno elencati nel nugget prodotto dal nodo. I modelli con la classificazione più alta vengono elencati in base al criterio di classificazione specificato. Si noti che l'incremento di questo limite potrebbe determinare una diminuzione delle performance. Il valore massimo consentito è 100.

Misura di classificazione silhouette

La misura di classificazione di default, Silhouette, ha un valore di default pari a 0 in quanto un valore inferiore a 0 (negativo) indica che la distanza media tra un caso e i punti nel cluster assegnato è superiore alla distanza media minima dai punti in un altro cluster. Pertanto, i modelli con una Silhouette negativa possono essere tranquillamente scartati.

La misura di classificazione è effettivamente un coefficiente di Silhouette modificato, che riunisce il concetto di coesione dei cluster (che predilige i modelli contenenti cluster strettamente coesi) e il concetto di separazione dei cluster (che predilige i modelli contenenti cluster molto separati). Il coefficiente medio di Silhouette è semplicemente la media calcolata su tutti i casi del seguente calcolo per ogni singolo caso:

$(B - A) / \max(A, B)$

dove A è la distanza dal caso al centroide del cluster a cui il caso appartiene e B è la distanza minima dal caso al centroide di ogni altro caso.

Il coefficiente di Silhouette (e la sua media) vanno da -1 (modello molto scarso) a 1 (modello eccellente). La media può essere calcolata sul livello dei casi totali (che produce la Silhouette totale) o sul livello di cluster (che produce la Silhouette di cluster). Le distanze possono essere calcolate mediante le distanze euclidee.

Opzioni avanzate del nodo Cluster automatico

La scheda Livello avanzato del nodo Cluster automatico consente di applicare una partizione (se disponibile), selezionare gli algoritmi da utilizzare e specificare le regole di arresto.

Modelli utilizzati. Utilizzare le caselle di controllo nella colonna di sinistra per selezionare i tipi di modelli (algoritmi) da includere nel confronto. Più tipi di modelli vengono selezionati, più modelli verranno creati e più lunghi saranno i tempi di elaborazione.

Tipo di modello. Elenca gli algoritmi disponibili (vedere di seguito).

Parametri del modello. Per ogni tipo di modello è possibile utilizzare le impostazioni di default oppure selezionare **Specifica** per scegliere le opzioni relative ai singoli tipi di modello. Le opzioni specifiche sono simili a quelle disponibili nei singoli nodi Modelli, con la differenza che è possibile selezionare più opzioni o combinazioni di opzioni. Per esempio, se si confrontano modelli Rete neurale, invece di scegliere uno dei sei metodi di addestramento, è possibile scegliere tutti i metodi in modo da addestrare sei modelli in un unico passaggio.

Numero di modelli. Elenca il numero di modelli prodotti per ciascun algoritmo in base alle impostazioni correnti. Quando si sceglie una combinazione di opzioni, il numero di modelli può aumentare rapidamente; pertanto, si consiglia di fare molta attenzione a questo numero, soprattutto se si utilizzano insiemi di dati molto estesi.

Limita tempo massimo impiegato per creare un modello singolo. (Solo modelli Medie K, Kohonen, TwoStep, SVM, KNN, Rete di Bayes ed Elenco di decisioni) Imposta un limite di tempo massimo per ogni modello. Per esempio, se un modello richiede un tempo di addestramento più lungo del previsto a causa di qualche interazione complessa, grazie a questa opzione è possibile evitare che l'operazione ritardi e blocchi l'esecuzione di tutta la modellazione.

Algoritmi supportati

Il nodo Medie K raggruppa l'insieme di dati in gruppi distinti (o cluster). Il metodo definisce un numero fisso di cluster, esegue un'assegnazione iterativa dei record ai cluster e modifica i centri di cluster finché un'ulteriore ridefinizione non consente più un miglioramento del modello. Invece di tentare di prevedere un risultato, il nodo *K*-medie utilizza un processo denominato apprendimento non supervisionato per scoprire gli schemi nell'insieme di campi di input.

Il nodo Kohonen genera un tipo di rete neurale che può essere utilizzato per raggruppare l'insieme di dati in gruppi distinti. Al termine dell'apprendimento della rete, i record analoghi dovranno essere vicini nella mappa di output, mentre i record diversi saranno a notevole distanza. Per identificare le unità forti, è possibile controllare il numero di osservazioni catturate da ciascuna unità nel nugget del modello. In questo modo è possibile avere un'idea del numero appropriato di cluster.

Il nodo TwoStep è un metodo di raggruppamento tramite cluster in due fasi. La prima fase esegue un singolo passaggio nei dati per comprimere i dati di input non elaborati in un insieme gestibile di cluster secondari. Nella seconda fase viene utilizzato un metodo di raggruppamento tramite cluster gerarchico per unire progressivamente i cluster secondari in cluster sempre più grandi. Il nodo TwoStep offre il vantaggio di stimare automaticamente il numero ottimale di cluster per i dati di addestramento. Può gestire in modo efficiente tipi di campo misti e insiemi di dati di grandi dimensioni.

Opzioni di scarto del nodo Cluster automatico

La scheda Scarta del nodo Cluster automatico consente di scartare automaticamente i modelli che non soddisfano determinati criteri. Questi modelli non verranno inclusi nel nugget del modello.

È possibile specificare il valore minimo di silhouette, i numeri di cluster, le dimensioni dei cluster e l'importanza del campo di valutazione usato nel modello. La silhouette e il numero e la dimensione dei cluster sono determinati nel modo specificato nel nodo Modelli. Per ulteriori informazioni, consultare l'argomento “Opzioni del modello di nodo Cluster automatico” a pagina 75.

Se lo si desidera, è possibile configurare il nodo in modo che l'esecuzione venga interrotta alla prima generazione di un modello che soddisfa tutti i criteri specificati. Per ulteriori informazioni, consultare l'argomento “Regole di arresto dei nodi Modelli automatici” a pagina 64.

Nugget del modello automatici

Quando viene eseguito un nodo Modelli automatici, il nodo stima i modelli candidati per ogni possibile combinazione di opzioni, classifica ogni modello candidato in base alla misura specificata dall'utente e salva i modelli migliori in un nugget del modello automatico composito. Il nugget del modello contiene in realtà una serie di uno o più modelli generati dal nodo, che è possibile consultare separatamente o selezionare per l'impiego nel calcolo del punteggio. Per ogni modello vengono indicati il tipo e il tempo di creazione, oltre a svariate altre misure a seconda del tipo di modello. La tabella può essere ordinata in base a qualsiasi colonna in modo da poter identificare rapidamente i modelli più interessanti.

- Per consultare uno dei singoli nugget del modello, fare doppio clic sull'icona dell'insieme. Da qui è possibile generare un nodo Modelli per quel modello nell'area del flusso o una copia del nugget del modello nella palette dei modelli.
- I grafici in miniatura consentono una rapida valutazione visiva di ogni modello, come riassunto di seguito. Per generare un grafico a schermo intero è possibile fare doppio clic su una miniatura. Il grafico a schermo intero riproduce un massimo di 1.000 punti e si baserà su un campione se l'insieme di dati ne contiene di più. Solo per i grafici a dispersione, il grafico viene rigenerato ogni volta che viene visualizzato, per cui tutte le modifiche ai dati a monte — ad esempio l'aggiornamento di un campione casuale o di una partizione se non è selezionata l'opzione **Imposta seed random** — possono essere riportate ogni volta che il grafico a dispersione viene nuovamente tracciato.
- Per visualizzare o nascondere colonne specifiche nella scheda Modelli o per cambiare la colonna in base alla quale ordinare la tabella, utilizzare la barra degli strumenti. Per modificare l'ordinamento è anche possibile fare clic sulle intestazioni delle colonne.
- Utilizzare il pulsante Elimina per eliminare in modo definitivo i modelli inutilizzati.
- Per riordinare le colonne, fare clic su un'intestazione e trascinare la colonna nella posizione desiderata.
- Se si utilizza una partizione, è possibile scegliere di visualizzare i risultati relativi alla partizione di addestramento o alla partizione di test, a seconda del caso.

Le colonne specifiche dipendono dai tipi di modelli oggetto del confronto, come illustrato di seguito.

Obiettivi binari

- Per i modelli binari, il grafico in miniatura mostra la distribuzione dei valori effettivi a cui vengono sovrapposti quelli previsti, per fornire una rapida indicazione visiva del numero di record che sono stati previsti correttamente in ogni categoria.
- I criteri di classificazione corrispondono alle opzioni del nodo Modelli Classificatore automatico. Per ulteriori informazioni, consultare l'argomento “Opzioni del modello di nodo Classificatore automatico” a pagina 65.
- Per il profitto massimo, viene riportato anche il percentile nel quale tale livello massimo si verifica.
- Per il guadagno cumulativo, è possibile utilizzare la barra degli strumenti per modificare il percentile selezionato.

Obiettivi nominali

- Per i modelli nominali (di insiemi), il grafico in miniatura mostra la distribuzione dei valori effettivi a cui vengono sovrapposti quelli previsti, per fornire una rapida indicazione visiva del numero di record che sono stati previsti correttamente in ogni categoria.

- I criteri di classificazione corrispondono alle opzioni del nodo Modelli Classificatore automatico. Per ulteriori informazioni, consultare l'argomento “Opzioni del modello di nodo Classificatore automatico” a pagina 65.

Target continui

- Per i modelli continui (di intervallo numerico), il grafico rappresenta i valori previsti rispetto a quelli osservati per ogni modello e fornisce una rapida indicazione visiva della loro correlazione. In un modello efficace, i punti tendono a raggrupparsi lungo la diagonale anziché essere distribuiti a caso in tutto il grafico.
- I criteri di classificazione corrispondono alle opzioni del nodo Modelli Numerico automatico. Per ulteriori informazioni, consultare l'argomento “Opzioni del modello di nodo Numerico automatico” a pagina 71.

Obiettivi cluster

- Per i modelli di cluster, il grafico rappresenta i conteggi rispetto ai cluster per ogni modello fornendo una rapida indicazione visiva della distribuzione dei cluster.
- I criteri di classificazione corrispondono alle opzioni del nodo Modelli Cluster automatico. Per ulteriori informazioni, consultare l'argomento “Opzioni del modello di nodo Cluster automatico” a pagina 75.

Selezione dei modelli per il calcolo del punteggio

La colonna **Uso?** consente di selezionare i modelli da usare per il calcolo del punteggio.

- Per gli obiettivi binari, nominali e numerici è possibile selezionare più modelli per il calcolo del punteggio e combinare i punteggi nei risultati di un unico nugget del modello di classificazione binario. Se si combinano le previsioni di più modelli è possibile superare le limitazioni dei singoli modelli e ottenere così un grado di precisione complessivo migliore rispetto a quello ottenibile da ciascun modello.
- Per i modelli di cluster è possibile selezionare un solo modello di calcolo del punteggio alla volta. Di default viene selezionato il primo classificato.

Generazione di nodi e modelli

È possibile generare una copia del nugget del modello automatico composito o il nodo modelli automatici da cui è stato creato. Questa operazione potrebbe essere utile, per esempio, se non si dispone del flusso originale da cui il nugget del modello automatico era stato creato. In alternativa, è possibile generare un nugget o un nodo Modelli per ciascuno dei modelli elencati nel nugget del modello automatico.

Nugget del modello automatico

Dal menu Genera, selezionare **Modello a palette** per aggiungere il nugget del modello automatico alla palette Modelli. Il modello generato può essere salvato o utilizzato così com'è, senza eseguire nuovamente il flusso.

In alternativa, è possibile selezionare **Genera nodo Modelli** dal menu Genera per aggiungere il nodo Modelli all'area del flusso. Questo nodo può essere utilizzato per stimare nuovamente i modelli selezionati senza che sia necessario ripetere tutta l'esecuzione della modellazione.

Nugget del modello singolo

1. Nel menu **Modelli**, fare doppio clic sul nugget singolo richiesto. Una copia del nugget viene aperta in una nuova finestra di dialogo.
2. Dal menu Genera della nuova finestra, selezionare **Modello a palette** per aggiungere il nugget del modello singolo alla palette Modelli.

3. In alternativa, è possibile selezionare **Genera nodo Modelli** dal menu Genera nella nuova finestra di dialogo per aggiungere il nodo Modelli singolo all'area del flusso.

Generazione di grafici di valutazione

Per i soli modelli binari è possibile generare grafici di valutazione che consentono di valutare visivamente e di confrontare le prestazioni di ogni modello. I grafici di valutazione non sono disponibili per i modelli generati dal nodo Numerico automatico o Cluster automatico.

1. Nella colonna *Uso?* del nugget del modello automatico Classificatore automatico, selezionare i modelli che si desidera valutare.
2. Dal menu Genera, scegliere **Grafico/i di valutazione**. Viene visualizzata la finestra di dialogo Grafico di valutazione.
3. Selezionare il tipo di grafico e specificare le opzioni desiderate.

Grafici di valutazione

Nella scheda Modelli del nugget del modello automatico, è possibile eseguire il drill-down per visualizzare i singoli grafici per ognuno dei modelli visualizzati. Nel caso di nugget Classificatore automatico e Numerico automatico, la scheda Grafico visualizza sia un grafico sia l'importanza dei predittori che riflettono i risultati di tutti i modelli combinati. Per ulteriori informazioni, consultare l'argomento “Importanza predittore” a pagina 43.

Nel caso del Classificatore automatico, viene visualizzato un grafico di distribuzione, mentre per il Numerico automatico viene visualizzato un multiplot (anche noto come grafico a dispersione).

Capitolo 6. Strutture ad albero delle decisioni

Modelli di struttura ad albero delle decisioni

I modelli della struttura ad albero delle decisioni consentono di sviluppare sistemi di classificazione in grado di prevedere o classificare osservazioni future in base ad un insieme di regole decisionali. Se i dati sono divisi in classi di interesse (per esempio, prestiti a basso vs. alto rischio, sottoscrittore vs. non sottoscrittore, votanti vs. non votanti, oppure tipi di batteri), è possibile utilizzare i dati per generare regole da utilizzare per classificare casi precedenti e nuovi con la massima precisione. Per esempio, è possibile creare una struttura ad albero che classifica il rischio sul credito o l'intento di acquisto in base all'età e ad altri fattori.

Tale approccio, anche noto come **induzione di regole**, ha vari vantaggi. Primo, il processo decisionale dietro al modello è evidente quando si sfoglia la struttura ad albero. Ciò contrasta con altre tecniche di modellazione di tipo "black box" in cui la logica interna è difficile da capire.

Secondo, il processo include automaticamente nella propria regola solo gli attributi realmente rilevanti nella decisione. Gli attributi irrilevanti ai fini della precisione della struttura ad albero verranno ignorati. Questo metodo è in grado di restituire informazioni estremamente importanti sui dati e può essere utilizzato per ridurre i dati includendo i campi pertinenti prima della preparazione di un'altra tecnica di apprendimento, quale quella delle reti neurali.

È possibile convertire nugget del modello di strutture ad albero delle decisioni generati in una raccolta di regole se-allora (un **insieme di regole**), che in molti casi mostra le informazioni in una forma più comprensibile. La presentazione di strutture ad albero delle decisioni si rivela utile quando si desidera visualizzare le **suddivisioni** o **partizioni** dei dati in sottoinsiemi pertinenti per il problema create dagli attributi all'interno dei dati. La presentazione di insiemi di regole si rivela utile quando si desidera visualizzare la relazione tra un determinato gruppo di elementi e una conclusione specifica. La regola riportata di seguito, per esempio, fornisce un **profilo** per un gruppo di automobili da acquistare.

```
IF tested = 'yes'
AND mileage = 'low'
THEN -> 'BUY'.
```

Algoritmi di creazione della struttura ad albero

Per l'esecuzione dell'analisi della segmentazione e della classificazione, sono disponibili quattro algoritmi. Fondamentalmente, tali algoritmi eseguono tutti la stessa operazione - esaminano tutti i campi del dataset per individuare il campo che fornisce la migliore classificazione o previsione suddividendo i dati in sottogruppi. Il processo viene applicato in modo ricorsivo, suddividendo i sottogruppi in unità sempre più piccole fino alla fine della struttura ad albero, secondo la definizione di alcuni criteri di interruzione. I campi obiettivo e di input utilizzati nella creazione di strutture ad albero possono essere continui (intervallo numerico) o categoriali, a seconda dell'algoritmo utilizzato. Se viene utilizzato un target continuo, viene generata una struttura ad albero di regressione; se viene utilizzato un obiettivo categoriale, viene generata una struttura ad albero di classificazione.

Il nodo Struttura ad albero di classificazione e regressione (C&R) genera una struttura ad albero delle decisioni che consente di prevedere o classificare osservazioni future. Il metodo utilizza partizionamento ricorsivo per suddividere i record di addestramento in segmenti, riducendo l'impurità ad ogni passaggio. Un nodo della struttura ad albero è considerato “puro” quando il 100% dei casi nel nodo fa parte di una categoria specifica del campo obiettivo. I campi obiettivo e di input possono essere intervalli numerici o categoriali (nominali, ordinali o flag); tutte le suddivisioni sono binarie (solo due sottogruppi).

Il nodo CHAID genera una struttura ad albero delle decisioni utilizzando statistiche chi-quadrato per identificare suddivisioni ottimali. A differenza dei nodi C&R Tree e QUEST, il nodo CHAID può generare strutture ad albero non binarie e pertanto alcune suddivisioni possono avere più di due rami. I campi obiettivo e di input possono essere intervallo numerico (continui) o categoriali. Un CHAID completo è una modificazione di CHAID che esegue operazioni avanzate per l'analisi di tutte le suddivisioni possibili, ma richiede tempi di elaborazione maggiori.

Il nodo QUEST offre un metodo di classificazione binario per la creazione di strutture ad albero delle decisioni, progettato per ridurre i tempi di elaborazione necessari per le analisi C&R Tree più complesse, riducendo inoltre la tendenza dei metodi per le strutture ad albero di classificazione a favorire gli input che consentono un numero maggiore di suddivisioni. I campi di input possono essere intervalli numerici (continui), ma il campo obiettivo deve essere categoriale. Tutte le suddivisioni sono binarie.

Il nodo C5.0 crea una struttura ad albero delle decisioni o un insieme di regole. Il modello suddivide il campione in base al campo che fornisce il massimo guadagno di informazioni a ogni livello. Il campo obiettivo deve essere categoriale. Sono consentite suddivisioni multiple in più di due sottogruppi.

Utilizzi generali dell'analisi basata su strutture ad albero

Di seguito sono riportati alcuni utilizzi generali dell'analisi basata su strutture ad albero:

Segmentazione. Identifica le persone che hanno più probabilità di essere membri di una particolare classe.

Stratificazione. Assegna i casi a più categorie, ad esempio gruppi ad alto, medio e basso rischio.

Previsione. Consente di creare regole e di utilizzarle per la previsione di eventi futuri. La previsione può anche significare il tentativo di mettere in relazione gli attributi predittivi con i valori di una variabile continua.

Riduzione dei dati ed esame delle variabili. Selezionare un sottoinsieme utile di predittori da un insieme di variabili di grandi dimensioni da utilizzare per la creazione di un modello parametrico formale.

Identificazione di interazioni. Identifica le relazioni relative solo a sottogruppi specifici e le specifica in un modello parametrico formale.

Unione di categorie e divisione in sezioni delle variabili continue. Ricodificare le categorie di predittori di gruppo e le variabili continue con una perdita minima di informazioni.

Builder della struttura ad albero interattiva

È possibile creare automaticamente un modello di struttura ad albero consentendo all'algoritmo di scegliere la suddivisione migliore per ogni livello oppure utilizzare il Builder della struttura ad albero interattiva e rifinire o semplificare la struttura ad albero applicando le informazioni di business note prima di salvare il nugget del modello.

1. Creare un flusso ed aggiungere uno dei nodi della struttura ad albero delle decisioni C&R Tree, CHAID o QUEST.

   Nota: la creazione della struttura ad albero interattiva non è supportata per le strutture ad albero C5.0.

2. Aprire il nodo e, nella scheda Campi, selezionare i campi obiettivo e predittore e specificare ulteriori opzioni di modelli secondo le proprie esigenze. Per istruzioni specifiche, consultare la documentazione di ogni nodo di creazione della struttura ad albero.
3. Nel riquadro Obiettivi della scheda Opzioni di creazione selezionare **Avvia sessione interattiva**.
4. Fare clic su **Esegui** per avviare il builder della struttura ad albero.

Viene visualizzata la struttura ad albero corrente, a partire dal nodo root. È possibile modificare e tagliare la struttura ad albero livello per livello e accedere ai guadagni, ai rischi e alle informazioni correlate prima di generare uno o più modelli.

Commenti

- Con i nodi C&R Tree, CHAID e QUEST, i campi ordinali utilizzati nel modello devono avere una classe di archiviazione numerica (non stringa). Se è necessario convertirli, è possibile utilizzare il nodo Ricodifica.
- È possibile utilizzare un campo partizione per suddividere i dati in campioni di addestramento e di test.
- In alternativa all'utilizzo del Builder della struttura ad albero, è inoltre possibile generare un modello direttamente dal nodo Modelli come per gli altri modelli di IBM SPSS Modeler. Per ulteriori informazioni, consultare l'argomento "Creazione diretta di un modello di struttura ad albero" a pagina 94.

Ingrandimento e taglio della struttura ad albero

La scheda Visualizzatore nel builder della struttura ad albero consente di visualizzare la struttura ad albero corrente, a partire dal nodo root.

1. Per ingrandire la struttura ad albero, dai menu scegliere:

   Struttura ad albero > **Ingrandisci struttura ad albero**

   Per creare la struttura ad albero, il sistema suddivide ogni ramo in modo ricorsivo fino a quando non viene soddisfatto un criterio di arresto. In corrispondenza di ogni suddivisione viene automaticamente selezionato il predittore migliore in base al metodo di modellazione utilizzato.
2. In alternativa, selezionare **Ingrandisci struttura ad albero di un livello** per aggiungere un singolo livello.
3. Per aggiungere un ramo sotto a un nodo specifico, selezionare il nodo e quindi **Ingrandisci ramo**.
4. Per scegliere il predittore da utilizzare per una suddivisione, selezionare il nodo desiderato e quindi **Ingrandisci ramo con suddivisione personalizzata**. Per ulteriori informazioni, consultare l'argomento "Definizione delle suddivisioni personalizzate" a pagina 84.
5. Per tagliare un ramo, selezionare un nodo e quindi scegliere **Rimuovi ramo** per cancellare il nodo selezionato.
6. Per rimuovere il livello inferiore della struttura ad albero, selezionare **Rimuovi di un livello**.
7. Solo per le strutture ad albero C&R e QUEST, selezionare **Ingrandisci struttura ad albero e taglia** per eseguire il taglio in base a un algoritmo di complessità del costo che modifica la stima del rischio in base al numero di nodi terminali e consente in genere di ottenere una struttura ad albero più semplice. Per ulteriori informazioni, consultare l'argomento "Nodo C&R Tree" a pagina 96.

Lettura delle regole di suddivisione nella scheda Visualizzatore

Quando si visualizzano le regole di suddivisione nella scheda Visualizzatore, le parentesi quadre indicano che il valore adiacente è incluso nell'intervallo, mentre le parentesi tonde indicano che il valore adiacente è escluso dall'intervallo. Pertanto, l'espressione (23,37] significa da 23 escluso a 37 incluso, cioè da appena sopra a 23 fino a 37. Nella scheda Modello, la stessa condizione sarebbe visualizzata come:

```
Age > 23 and Age <= 37
```

Interruzione dell'espansione della struttura ad albero. Per interrompere un'operazione di espansione della struttura ad albero, per esempio nel caso in cui tale operazione richieda più tempo del previsto, fare clic sul pulsante Interrompi esecuzione sulla barra degli strumenti.

Figura 28. Pulsante Interrompi esecuzione

Il pulsante è attivo solo durante l'espansione della struttura ad albero. Il processo di espansione viene interrotto immediatamente e i nodi aggiunti vengono mantenuti senza salvare le modifiche o chiudere la finestra. Il builder della struttura ad albero resta aperto, consentendo di generare un modello, aggiornare le direttive o esportare l'output nel formato appropriato, in base alle necessità.

Definizione delle suddivisioni personalizzate

La finestra di dialogo Definisci suddivisione consente di selezionare il predittore e di specificare le condizioni per ciascuna suddivisione.

1. Nel builder della struttura ad albero, selezionare un nodo nella scheda Visualizzatore e dai menu scegliere:

   Struttura ad albero > **Ingrandisci ramo con suddivisione personalizzata**
2. Selezionare il predittore desiderato dall'elenco a discesa oppure fare clic sul pulsante **Predittori** per visualizzare i dettagli relativi a ogni predittore. Per ulteriori informazioni, consultare l'argomento "Visualizzazione dei dettagli dei predittori".
3. È possibile accettare le condizioni di default per ogni suddivisione oppure selezionare**Personalizzata** per definire le condizioni desiderate.
   - Per i predittori continui (intervallo numerico), è possibile utilizzare i campi **Modifica valori di intervallo** per specificare l'intervallo dei valori che rientrano in ogni nuovo nodo.
   - Per i predittori categoriali, è possibile utilizzare i campi **Modifica valori di insieme** o **Modifica valori ordinali** per indicare valori specifici, o un intervallo di valori nel caso di un predittore ordinale, mappati a ogni nuovo nodo.
4. Selezionare **Ingrandisci** per ingrandire di nuovo il ramo utilizzando il predittore selezionato.

È in genere possibile suddividere la struttura ad albero con qualsiasi predittore, indipendentemente dalle regole di arresto. Le uniche eccezioni si verificano se il nodo è puro (ovvero quando il 100% dei casi rientra nella stessa classe di destinazione e pertanto non rimane nulla da suddividere) oppure se il predittore selezionato è una costante (non esiste nulla da suddividere).

Valori mancanti in. Per le strutture ad albero CHAID, se per un predittore sono presenti valori mancanti, quando si definisce una suddivisione personalizzata è possibile assegnarle a un nodo figlio specifico. Con C&R Tree e QUEST, i valori mancanti vengono gestiti utilizzando i surrogati definiti nell'algoritmo. Per ulteriori informazioni, consultare l'argomento "Dettagli e surrogati delle suddivisioni" a pagina 85.

Visualizzazione dei dettagli dei predittori

Nella finestra di dialogo Seleziona predittore vengono visualizzate le statistiche relative ai predittori disponibili (a volte denominati "concorrenti") che possono essere utilizzate per la suddivisione corrente.

- Per le strutture ad albero CHAID e CHAID completo, viene visualizzata la statistica chi-quadrato per ogni predittore categoriale. Se il predittore è un intervallo numerico, viene visualizzata la statistica *F*. La statistica del chi-quadrato è una misura del grado di indipendenza del campo obiettivo rispetto al campo di suddivisione. Generalmente una statistica del chi-quadrato elevata fa riferimento ad una probabilità più bassa; ciò significa che esistono meno possibilità che i due campi siano indipendenti — un'indicazione che la suddivisione è corretta. Sono inclusi anche i gradi di libertà in quanto è più facile per una suddivisione a tre vie avere un'ampia statistica e una probabilità bassa rispetto a una suddivisione a due vie.

- Per C&R Tree e QUEST, viene visualizzato il miglioramento per ciascun predittore. Maggiore è il miglioramento, maggiore è la riduzione dell'impurità tra i nodi padre e figlio se viene utilizzato tale predittore. Un nodo puro è un nodo nel quale tutti i casi rientrano in una singola categoria obiettivo, pertanto minore è l'impurità nella struttura ad albero e maggiore sarà la corrispondenza del modello ai dati. In altre parole, una figura con un elevato miglioramento in genere indica una suddivisione utile per questo tipo di struttura ad albero. La misura di impurità utilizzata è specificata nel nodo di creazione della struttura ad albero.

Dettagli e surrogati delle suddivisioni

Per visualizzare i dettagli relativi alla suddivisione per un nodo, è possibile selezionare il nodo nella scheda Visualizzatore e quindi fare clic sul pulsante Info suddivisione a destra sulla barra degli strumenti. Verranno visualizzate la regola di suddivisione e le relative statistiche. Per le strutture ad albero categoriali C&R Tree, sono visualizzati il miglioramento e l'associazione. L'associazione è una misura di corrispondenza tra un surrogato ed il campo di suddivisione principale, con il surrogato "migliore" che generalmente è quello che più assomiglia al campo di suddivisione. Per C&R Tree e QUEST, sono elencati anche i surrogati utilizzati al posto del predittore principale.

Per modificare la suddivisione per il nodo selezionato, è possibile fare clic sull'icona a sinistra del riquadro dei surrogati per aprire la finestra di dialogo Definisci suddivisione. In alternativa, è possibile selezionare un surrogato dall'elenco prima di fare clic sull'icona per classificarlo come campo di suddivisione principale.

Surrogati. Se possibile, qualsiasi surrogato per il campo della suddivisione principale viene visualizzato per il nodo selezionato. I surrogati sono campi alternativi utilizzati quando manca il valore del predittore principale per un record determinato. Il numero massimo di surrogati consentito per una determinata suddivisione viene specificato nel nodo di creazione della struttura ad albero ma il numero effettivo dipende dai dati di addestramento. In generale, più dati mancano, più surrogati vengono utilizzati. In altri modelli di strutture ad albero delle decisioni, questa scheda è vuota.

Nota: per essere inclusi nel modello, i surrogati devono essere identificati durante la fase di addestramento. Se il campione di addestramento non ha valori mancanti, nessun surrogato verrà identificato e qualsiasi record con valori mancanti trovato durante i test o il calcolo del punteggio rientra automaticamente nel nodo figlio con il numero maggiore di record. Se sono previsti valori mancanti durante i test o il calcolo del punteggio, accertarsi che i valori siano assenti anche nel campione di addestramento. I surrogati non sono disponibili per le strutture ad albero CHAID.

I surrogati non vengono utilizzati per le strutture ad albero CHAID, ma quando si definisce una suddivisione personalizzata è possibile assegnarle a un nodo figlio specifico. Per ulteriori informazioni, consultare l'argomento "Definizione delle suddivisioni personalizzate" a pagina 84.

Personalizzazione della vista della struttura ad albero

Nella scheda Visualizzatore del Builder della struttura ad albero viene visualizzata la struttura ad albero corrente. Tutti i rami della struttura ad albero sono espansi per default, ma è possibile espandere e comprimere i rami e personalizzare le altre impostazioni in base alle specifiche esigenze.

- Fare clic sul segno meno (–) nell'angolo in basso a destra di un nodo padre per nascondere tutti i relativi nodi figlio. Fare clic sul segno più (+) nell'angolo inferiore destro di un nodo padre per visualizzare tutti i nodi figlio.
- Utilizzare il menu o la barra degli strumenti Visualizza per modificare l'orientamento della struttura ad albero (dall'alto in basso, da sinistra a destra o da destra a sinistra).
- Fare clic sul pulsante "Visualizza le etichette di valori e campi" della barra degli strumenti principale per visualizzare o nascondere le etichette di campi e valori.
- Utilizzare i pulsanti con la lente di ingrandimento per eseguire uno zoom avanti o indietro oppure fare clic sul pulsante Mappa della struttura ad albero a destra sulla barra degli strumenti per visualizzare un diagramma dell'intera struttura ad albero.

- Se si utilizza un campo partizione, è possibile passare dalla vista della struttura ad albero della partizione di addestramento a quella della partizione di test e viceversa (**Visualizza > Partizione**). Quando è visualizzato il campione di test, la struttura ad albero può essere visualizzata ma non modificata. La partizione corrente viene visualizzata nella barra di stato disponibile nell'angolo in basso a destra della finestra.
- Fare clic sul pulsante di informazioni sulla suddivisione (il pulsante "i" all'estremità destra della barra degli strumenti) per visualizzare i dettagli della suddivisione corrente. Per ulteriori informazioni, consultare l'argomento “Dettagli e surrogati delle suddivisioni” a pagina 85.
- Per ogni nodo, è possibile visualizzare le statistiche o i grafici oppure entrambi (vedere sezione seguente).

Visualizzazione delle statistiche e dei grafici

Statistiche del nodo. Per un campo obiettivo categoriale, la tabella di ogni nodo mostra il numero e la percentuale di record in ogni categoria e la percentuale dell'intero campione rappresentata dal nodo. In un campo target continuo (intervallo numerico), la tabella mostra la media, la deviazione standard, il numero di record e il valore previsto del campo obiettivo.

Grafici del nodo. Per un campo obiettivo categoriale, il grafico è un grafico a barre che rappresenta le percentuali in ogni categoria del campo obiettivo. Ogni riga nella tabella è preceduta da un quadratino colorato corrispondente al colore che rappresenta ognuna delle categorie del campo obiettivo nei grafici relativi al nodo. Nel campo target continuo (intervallo numerico), il grafico visualizza un istogramma del campo obiettivo relativo ai record nel nodo.

Guadagni

La scheda Guadagni consente di visualizzare statistiche per tutti i nodi terminali nella struttura ad albero. I guadagni forniscono una misura della differenza tra la media o la proporzione in un nodo specifico e la media globale. In generale, maggiore è la differenza e più utile risulterà la struttura ad albero come strumento decisionale. Per esempio, un indice o valore guadagno cumulativo del 148% per un nodo indica che le probabilità che i record del nodo rientrino nella categoria obiettivo sono pari a una volta e mezza quelle dell'intero insieme di dati.

Per i nodi C&R Tree e QUEST in cui è specificato un insieme di prevenzione del sovradattamento vengono visualizzati due insiemi di statistiche:

- insieme di espansione della struttura ad albero - il campione di addestramento da cui è stato eliminato l'insieme di prevenzione del sovradattamento
- insieme di prevenzione del sovradattamento

Per le altre strutture ad albero interattive C&R Tree e QUEST e per tutte le altre strutture ad albero interattive CHAID sono visualizzate solo le statistiche dell'insieme di espansione della struttura ad albero.

La scheda Guadagni consente di:

- Visualizzare le statistiche nodo per nodo, le statistiche cumulative o i quantili.
- Visualizzare i guadagni o i profitti.
- Passare dalla visualizzazione delle tabelle a quella dei grafici e viceversa.
- Selezionare la categoria obiettivo (solo obiettivi categoriali).
- Ordinare la tabella in ordine crescente o decrescente in base alla percentuale dell'indice. Se sono visualizzate le statistiche per più partizioni, gli ordinamenti vengono sempre applicati al campione di addestramento anziché al campione di test.

In generale, le selezioni eseguite nella tabella dei guadagni verranno aggiornate nella vista della struttura ad albero e viceversa. Per esempio, se si seleziona una riga della tabella, nella struttura ad albero verrà selezionato il nodo corrispondente.

Guadagni di classificazione

Per le strutture ad albero di classificazione (strutture ad albero con una variabile obiettivo categoriale), la percentuale dell'indice cumulato dei guadagni indica la differenza tra la proporzione di una categoria obiettivo specifica in ogni nodo e la proporzione globale.

Statistiche Nodo per nodo

In questa visualizzazione, la tabella mostra una riga per ogni nodo terminale. Per esempio, se la risposta globale a una campagna per posta è stata del 10%, ma il 20% dei record contenuti nel nodo X ha risposto favorevolmente, la percentuale dell'indice cumulato del nodo sarà 200%, a indicare che la probabilità che i rispondenti di questo gruppo effettuino acquisti è doppia rispetto a quella della popolazione globale.

Per i nodi C&R Tree e QUEST in cui è specificato un insieme di prevenzione del sovradattamento vengono visualizzati due insiemi di statistiche:

- insieme di espansione della struttura ad albero - il campione di addestramento da cui è stato eliminato l'insieme di prevenzione del sovradattamento
- insieme di prevenzione del sovradattamento

Per le altre strutture ad albero interattive C&R Tree e QUEST e per tutte le altre strutture ad albero interattive CHAID sono visualizzate solo le statistiche dell'insieme di espansione della struttura ad albero.

Nodi. ID del nodo corrente, come viene visualizzato nella scheda Visualizzatore.

Nodo: n. Il numero totale di record in tale nodo.

Nodo (%). Percentuale di tutti i record dell'insieme di dati contenuti nel nodo.

Guadagno: n. Il numero di record con la categoria obiettivo selezionata che rientrano in questo nodo. Corrisponde al numero dei record dell'insieme di dati che rientrano nella categoria obiettivo e che sono presenti nel nodo.

Guadagno (%). Percentuale di tutti i record nella categoria obiettivo e di tutto l'insieme di dati che sono contenuti nel nodo.

Risposta (%). Percentuale dei record del nodo corrente che sono contenuti nella categoria obiettivo. Le risposte in questo contesto vengono a volte denominate "risultati".

Indice (%). Percentuale di risposta per il nodo corrente espressa come percentuale della risposta percentuale per l'intero insieme di dati. Per esempio, un valore indice del 300% indica che la probabilità che i record di questo nodo siano contenuti nella categoria obiettivo è tre volte maggiore della probabilità dell'intero insieme di dati.

Statistiche cumulative

Nella visualizzazione cumulata, la tabella mostra un nodo per riga, ma le statistiche sono cumulative, ordinate in ordine crescente o decrescente in base alla percentuale dell'indice. Per esempio, se viene applicato un ordinamento decrescente, verrà visualizzato per primo il nodo con la percentuale dell'indice cumulato più alta e le statistiche nelle righe successive sono cumulative per la riga specifica e per quella superiore.

La percentuale dell'indice cumulato diminuisce riga per riga se vengono aggiunti nodi con risposte percentuali sempre più basse. L'indice cumulato per la riga finale è sempre il 100%, perché a questo punto è incluso l'intero insieme di dati.

Quantili

In questa visualizzazione, ogni riga della tabella rappresenta un quantile anziché un nodo. I quantili sono quartili, quintili (quinti), decili (decimi), ventili (ventesimi) o percentili (centesimi). È possibile che in un singolo quantile siano elencati più nodi, se per raggiungere la percentuale sono necessari più nodi, per esempio se sono visualizzati quartili ma i due nodi superiori contengono meno del 50% di tutti i casi. La parte rimanente della tabella è cumulata e può essere interpretata nello stesso modo della visualizzazione cumulata.

Profitti e ROI di classificazione

Per le strutture ad albero di classificazione, le statistiche dei guadagni possono essere visualizzate anche in termini di profitto e ROI (Return On Investment, ritorno dell'investimento). La finestra di dialogo Definisci profitti consente di specificare le entrate e le spese per ciascuna categoria.

1. Per accedere alla finestra di dialogo, fare clic sul pulsante Profitto (con etichetta $/$) nella scheda Guadagni.
2. Immettere i valori delle entrate e delle spese per ogni categoria del campo obiettivo.

Ad esempio, se il il costo dell'invio per posta di un'offerta a ciascun cliente è uguale a 0,48 euro e l'entrata per una risposta positiva è uguale a 9,95 euro per una sottoscrizione di tre mesi, il costo di ciascuna risposta *negativa* è 0,48 euro ed li guadagno per ciascuna risposta *positiva* corrisponde a 9,47 euro (9,95–0,48).

Nella tabella dei guadagni, il calcolo del **profitto** corrisponde alla somma delle entrate meno le spese per ognuno dei record in un nodo terminale. **ROI** corrisponde al profitto totale diviso per la spesa totale in un nodo.

Commenti

- I valori del profitto influiscono unicamente sui valori del profitto medio e sui valori ROI visualizzati nella tabella dei guadagni, rendendo le statistiche più comprensibili in termini di utile netto. Non influenzano nemmeno la struttura del modello della struttura ad albero di base. I profitti non devono essere confusi con i costi di errata classificazione, che sono specificati nel nodo di creazione della struttura ad albero e vengono scomposti in fattori nel modello allo scopo di fornire una protezione contro gli errori.
- Le specifiche dei profitti non vengono mantenute da una sessione di creazione della struttura ad albero interattiva a quella successiva.

Guadagni di regressione

Per le strutture ad albero di regressione, è possibile scegliere la visualizzazione nodo per nodo, nodo per nodo cumulata e dei quantili. Nella tabella sono visualizzati i valori medi. I grafici sono disponibili unicamente per i quantili.

Grafici dei profitti

In alternativa alle tabelle, nella scheda Guadagni è possibile visualizzare grafici.

1. Nella scheda Guadagni, selezionare l'icona Quantili (la terza da sinistra sulla barra degli strumenti). I grafici non sono disponibili per le statistiche nodo per nodo o cumulative.
2. Selezionare l'icona dei grafici.
3. Selezionare le unità da visualizzare (percentili, decili e così via) dalla casella di riepilogo.
4. Selezionare **Guadagni**, **Risposta** o **Guadagno cumulativo** per modificare la misura visualizzata.

Grafico dei profitti

Nel grafico dei profitti sono rappresentati i valori della colonna *Guadagni (%)* della tabella. I guadagni sono definiti come la proporzione di risultati in ogni incremento rispetto al numero totale di risultati nella struttura ad albero, mediante l'equazione seguente:

(risultati in incremento /numero totale di risultati) x 100%

Il grafico illustra la modalità di progettazione più idonea per identificare una percentuale specifica di tutti i risultati nella struttura ad albero. La linea diagonale rappresenta la risposta prevista per l'intero campione, se il modello non viene utilizzato. In questo caso, la percentuale di risposta è costante, poiché la probabilità che un utente risponda è uguale a quella di un altro utente. Per raddoppiare le possibilità, sarebbe necessario formulare la domanda a un numero doppio di persone. La linea curva indica di quanto è possibile aumentare la risposta inserendo solo le persone classificate nei percentili più alti in base al guadagno. Per esempio, se si include il 50% superiore, si può ottenere un incremento netto del 70 percento di risposte positive. Maggiore è la curvatura, maggiore è il guadagno.

Grafici guadagno cumulativo

Il grafico guadagno cumulativo rappresenta i valori della colonna *Indice (%)* della tabella. Il grafico confronta la percentuale di record che in ogni incremento corrispondono a risultati con la percentuale globale di risultati nell'insieme di dati di addestramento, mediante l'equazione seguente:

(risultati in incremento / record in incremento) / (numero totale di risultati /numero totale di record)

Grafico delle risposte

Il grafico delle risposte rappresenta i valori della colonna *Risposta (%)* della tabella. La risposta è una percentuale di record che nell'incremento corrispondono a risultati, in base all'equazione seguente:

(risposte in incremento / record in incremento) x 100%

Selezione in base a guadagni

La finestra di dialogo Selezione in base a guadagni consente di selezionare automaticamente i nodi terminale con il migliore (o peggiore) guadagno in base ad una soglia o una regola specificata. È quindi possibile generare un nodo Seleziona.

1. Nella scheda Guadagni, selezionare la visualizzazione nodo per nodo o cumulata e quindi la categoria obiettivo in base alla quale si desidera eseguire la selezione. Le selezioni sono basate sulla visualizzazione corrente della tabella e non sono disponibili per i quantili.
2. Nella scheda Guadagni, dai menu scegliere:

   Modifica > **Seleziona nodi terminali** > **Selezione in base a guadagni**

   Seleziona solo. È possibile selezionare nodi corrispondenti *o* nodi non corrispondenti — ad esempio, è possibile selezionare tutti i record *ad eccezione* dei primi 100.

   Confronta in base a informazioni sui guadagni. Trova le corrispondenze di nodi in base alle statistiche dei guadagni per la categoria obiettivo corrente, per esempio:
   - Nodi in cui il guadagno, la risposta o il guadagno cumulativo corrispondono ad una soglia specificata — ad esempio, risposta maggiore o uguale al 50%.
   - Primi *n* nodi in base al guadagno per la categoria obiettivo.
   - Primi nodi fino a un numero di record specificato.
   - Primi nodi fino a una percentuale di dati di addestramento specificata.
3. Fare clic su **OK** per aggiornare la selezione nella scheda Visualizzatore.
4. Per creare un nuovo nodo Seleziona in base alla selezione corrente nella scheda Visualizzatore, scegliere **Nodo Seleziona** dal menu Genera. Per ulteriori informazioni, consultare l'argomento “Generazione di nodi Filtro e Seleziona” a pagina 93.

Nota: la corrispondenza con il criterio di selezione potrebbe non essere sempre perfetta perché in effetti vengono selezionati nodi e non record o percentuali. Il sistema seleziona nodi completi *fino al* livello specificato. Per esempio, se si selezionano i primi 12 casi e 10 di essi sono nel primo nodo mentre gli altri due sono nel secondo nodo, verrà selezionato solo il primo nodo.

Rischi

I rischi indicano le probabilità di errata classificazione in qualsiasi livello. Nella tabella Rischi è visualizzata la stima puntuale del rischio e (per gli output categoriali) una tabella di errata classificazione.

- Per le previsioni numeriche, il rischio è rappresentato da una stima raggruppata della varianza in corrispondenza di ognuno dei nodi terminali.
- Per le previsioni categoriali, il rischio è rappresentato dalla proporzione dei casi classificati in modo non corretto, dei casi adeguati per le probabilità a priori o dei costi di errata classificazione.

Salvataggio di modelli di strutture ad albero delle decisioni e risultati

Esistono molti modi per salvare o esportare i risultati di una sessione di creazione della struttura ad albero interattiva, per esempio:

- Generare un modello in base alla struttura ad albero corrente (**Genera > Genera modello**).
- Salvare le direttive utilizzate per ingrandire la struttura ad albero corrente. Alla successiva esecuzione del nodo di creazione della struttura ad albero, la struttura ad albero corrente verrà di nuovo ingrandita automaticamente, incluse le eventuali suddivisioni personalizzate definite in precedenza.
- Esportare le informazioni relative al modello, al guadagno e ai rischi. Per ulteriori informazioni, consultare l'argomento "Esportare le informazioni relative al modello, al guadagno e ai rischi." a pagina 92.

Dal builder della struttura ad albero o da un nugget del modello della struttura ad albero, è possibile:

- Generare un nodo Filtro o Seleziona in base alla struttura ad albero corrente. Per ulteriori informazioni, consultare l'argomento "Generazione di nodi Filtro e Seleziona" a pagina 93.
- Generare un nugget del modello Insieme di regole contenente la struttura ad albero come insieme di regole che definiscono i rami finali della struttura ad albero. Per ulteriori informazioni, consultare l'argomento "Generazione di un Insieme di regole da una struttura ad albero delle decisioni" a pagina 93.
- Inoltre, solo per i nugget del modello della struttura ad albero, è possibile esportare il modello in formato PMML. Per ulteriori informazioni, consultare l'argomento "La palette Modelli" a pagina 40. Se il modello include suddivisioni personalizzate, tali informazioni non vengono conservate nel PMML esportato. (La suddivisione viene conservata, ma non il fatto che sia personalizzata anziché scelta dall'algoritmo).
- Generare un grafico in base a una parte selezionata della struttura ad albero corrente. *Nota*: questa operazione è valida per un nugget solo quando è collegato ad altri nodi in un flusso. Per ulteriori informazioni, consultare l'argomento "Generazione di grafici" a pagina 113.

Nota: non è possibile salvare la struttura ad albero interattiva. Per evitare di perdere il proprio lavoro, generare un modello e/o aggiornare le direttive della struttura ad albero prima di chiudere la finestra del Builder della struttura ad albero.

Generazione di un modello dal Builder della struttura ad albero

Per generare un modello in base alla struttura ad albero corrente, dai menu del Builder della struttura ad albero scegliere:

Genera > **Modello**

Nella finestra di dialogo Genera nuovo modello, è possibile scegliere tra le opzioni riportate di seguito:

Nome modello. È possibile specificare un nome personalizzato oppure generare automaticamente il nome in base al nome del nodo Modelli.

Crea nodo in. È possibile aggiungere il nodo all'**area**, alla **palette MG** o a **entrambe**.

Includi direttive della struttura ad albero. Selezionare questa casella per includere nel modello generato le direttive della struttura ad albero corrente. Ciò consente di rigenerare la struttura ad albero, se necessario. Per ulteriori informazioni, consultare l'argomento “Direttive di ingrandimento della struttura ad albero”.

Direttive di ingrandimento della struttura ad albero

Per i modelli C&R Tree, CHAID e QUEST, le direttive della struttura ad albero specificano le condizioni per la crescita della struttura ad albero, un livello alla volta. Le direttive vengono applicate ogni volta che si avvia il builder della struttura ad albero interattiva dal nodo.

- Le direttive risultano particolarmente utili per rigenerare una struttura ad albero creata durante una sessione interattiva precedente. Per ulteriori informazioni, consultare l'argomento “Aggiornamento delle direttive della struttura ad albero” a pagina 92. È inoltre possibile modificare le direttive manualmente, ma questa operazione deve essere eseguita con estrema attenzione.
- Le direttive sono altamente specifiche della struttura della struttura ad albero da esse definito, pertanto qualsiasi modifica ai dati o alle opzioni di modellazione sottostanti può determinare la mancata esecuzione di un insieme di direttive che in precedenza risultavano valide. Per esempio, se l'algoritmo CHAID modifica una suddivisione bidirezionale in una suddivisione tridirezionale in base ai dati aggiornati, qualsiasi direttiva basata sulla suddivisione precedente non verrà eseguita.

Nota: se si sceglie di generare un modello direttamente (senza utilizzare il builder della struttura ad albero), le direttive della struttura ad albero verranno ignorate.

Modifica di direttive

1. Per visualizzare o modificare le direttive salvate, aprire il nodo di creazione della struttura ad albero e selezionare il riquadro Obiettivo della scheda Opzioni di creazione.
2. Selezionare **Avvia sessione interattiva** per attivare i controlli, selezionare **Utilizza direttive della struttura ad albero** e quindi **Direttive**.

Sintassi delle direttive

Le direttive specificano le condizioni per l'ingrandimento della struttura ad albero, a partire dal nodo root. Per esempio, per ingrandire la struttura ad albero di un livello per volta:

```
Grow Node Index 0 Children 1 2
```

Poiché non viene specificato alcun predittore, l'algoritmo sceglie la suddivisione migliore.

Si noti che la prima suddivisione deve sempre essere nel nodo root (`Index 0`) ed è necessario specificare i valori indice per entrambi gli elementi figlio (`1` e `2` in questo caso). Non è corretto specificare `Grow Node Index 2 Children 3 4`, a meno che non sia stata prima ingrandita la radice che ha creato Node 2.

Per ingrandire la struttura ad albero:

```
Grow Tree
```

Per ingrandire e tagliare la struttura ad albero (solo C&R Tree):

```
Grow_And_Prune Tree
```

Per definire una suddivisione personalizzata per un predittore continuo:

```
Grow Node Index 0 Children 1 2 Spliton
  ( "EDUCATE", Interval ( NegativeInfinity, 12.5)
     Interval ( 12.5, Infinity ))
```

Per eseguire la suddivisione per un predittore nominale con due valori:

```
Grow Node Index 2 Children 3 4 Spliton
  ( "GENDER", Group( "0.0" )Group( "1.0" ))
```

Per eseguire la suddivisione per un predittore nominale con più valori:

```
Grow Node Index 6 Children 7 8 Spliton
   ( "ORGS", Group( "2.0","4.0" )
             Group( "0.0","1.0","3.0","6.0" ))
```

Per eseguire la suddivisione per un predittore ordinale:

```
Grow Node Index 4 Children 5 6 Spliton
  ( "CHILDS", Interval ( NegativeInfinity, 1.0)
              Interval ( 1.0, Infinity ))
```

Nota: quando si definiscono suddivisioni personalizzate, i nomi e i valori dei campi (`EDUCATE`, `GENDER`, `CHILDS`, ecc.) rilevano la distinzione tra caratteri maiuscoli/minuscoli.

Direttive per le strutture ad albero CHAID

Le direttive per le strutture ad albero CHAID sono particolarmente sensibili alle modifiche apportate ai dati o al modello perché, a differenza di C&R Tree e QUEST, non sono vincolate all'utilizzo delle suddivisioni binarie. Per esempio, la sintassi seguente sembra corretta, ma non può funzionare se l'algoritmo suddivide il nodo root in più di due elementi figlio:

```
Grow Node Index 0 Children 1 2
Grow Node Index 1 Children 3 4
```

Con CHAID, è possibile che il nodo 0 abbia 3 o 4 figli e pertanto la seconda riga della sintassi non verrà eseguita.

Utilizzo delle direttive negli script

Le direttive possono inoltre essere incorporate negli script mediante virgolette triple.

Aggiornamento delle direttive della struttura ad albero

Per mantenere il lavoro svolto durante una sessione di creazione di una struttura ad albero interattiva, è possibile salvare le direttive utilizzate per la generazione della struttura ad albero. A differenza del salvataggio di un nugget del modello, che non può essere ulteriormente modificato, in questo modo è possibile generare nuovamente la struttura ad albero nello stato corrente per ulteriori modifiche.

Per aggiornare le direttive, dai menu del builder della struttura ad albero scegliere:

File > **Aggiorna le direttive**

Le direttive vengono salvate nel nodo di modellazione utilizzato per creare la struttura ad albero (C&R Tree, QUEST o CHAID) e possono essere utilizzate per rigenerare la struttura ad albero corrente. Per ulteriori informazioni, consultare l'argomento “Direttive di ingrandimento della struttura ad albero” a pagina 91.

Esportare le informazioni relative al modello, al guadagno e ai rischi.

Dal builder della struttura ad albero è possibile esportare le statistiche del modello, del guadagno e del rischio in formato di testo, HTML o di immagine, in base alle specifiche esigenze.

1. Nella finestra del builder della struttura ad albero, selezionare la scheda o la visualizzazione che si desidera esportare.
2. Dai menu, scegliere:

   File > **Esporta**

3. Selezionare **Testo**, **HTML** o **Grafico** e quindi scegliere dal sottomenu gli elementi specifici che si desidera esportare.

Se possibile, l'esportazione sarà basata sulle selezioni correnti.

Esportazione in formato Testo o HTML. È possibile esportare le statistiche del guadagno o del rischio per la partizione di addestramento o di test (se definite). L'esportazione è basata sulle selezioni correnti nella scheda Guadagni — ad esempio, è possibile scegliere le statistiche del quantile, cumulative o nodo per nodo.

Esportazione di grafici. È possibile esportare la struttura ad albero corrente come viene visualizzata nella scheda Visualizzatore oppure esportare i grafici dei profitti per la partizione di addestramento o di test (se definite). I formati disponibili sono *.JPEG*, *.PNG* e *.BMP*. Per i guadagni, l'esportazione è basata sulle selezioni correnti nella scheda Guadagni (disponibile unicamente se è visualizzato un grafico).

Generazione di nodi Filtro e Seleziona

Nella finestra del builder della struttura ad albero oppure quando si sfoglia un nugget del modello della struttura ad albero delle decisioni, dai menu scegliere:

Generate > **Nodo Filtro**

oppure

> **Nodo Seleziona**

Nodo filtro. Genera un nodo che filtra i campi non utilizzati dalla struttura ad albero corrente. Ciò consente di ridurre in modo rapido l'insieme di dati, in modo da includervi unicamente i campi definiti come importanti dall'algoritmo. Se a monte del nodo Struttura ad albero delle decisioni esiste un nodo Tipo, qualsiasi campo con il ruolo *Obiettivo* viene passato dal nugget del modello Filtro.

Nodo Seleziona. Genera un nodo che seleziona tutti i record contenuti nel nodo corrente. Questa opzione richiede che nella scheda Visualizzatore siano selezionati uno o più rami della struttura ad albero.

Il nugget del modello viene collocato nell'area del flusso.

Generazione di un Insieme di regole da una struttura ad albero delle decisioni

È possibile generare un nugget del modello Insieme di regole che rappresenta la struttura ad albero come un insieme di regole che definiscono i rami terminali della struttura ad albero. Gli insiemi di regole spesso sono in grado di mantenere le informazioni più importanti di un'intera struttura ad albero delle decisioni ma con un modello meno complesso. La differenza più importante è il fatto che, con un insieme di regole, a un particolare record può applicarsi più di una regola o nessuna. Per esempio, è possibile che vengano visualizzate tutte le regole che prevedono un risultato *no*, seguite da tutte le regole che prevedono un risultato *sì*. Se si applicano più regole, ognuna di esse riceve un "voto" ponderato in base alla confidenza associata a quella regola e la previsione finale viene decisa combinando i voti ponderati di tutte le regole che si applicano al record interessato. Se non si applica alcuna regola, al record viene assegnata una previsione di default.

Gli insiemi di regole possono essere generati unicamente da strutture ad albero con campi obiettivo categoriali (non da strutture ad albero di regressione).

Nella finestra del builder della struttura ad albero oppure quando si sfoglia un nugget del modello della struttura ad albero delle decisioni, dai menu scegliere:

Genera > **Insieme di regole**

Nome insieme di regole. Consente di specificare il nome del nuovo nugget del modello Insieme di regole.

Crea nodo in. Controlla la posizione del nuovo nugget del modello Insieme di regole. Selezionare **Area**, **Palette GM** o **Entrambi**.

Istanze minime. Specifica il numero minimo di istanze (numero di record a cui si applica la regola) per mantenere il nugget del modello Insieme di regole. Le regole con un supporto inferiore al valore specificato non verranno incluse nel nuovo insieme di regole.

Confidenza minima. Specifica la confidenza minima per mantenere le regole nel nugget del modello Insieme di regole. Le regole con una confidenza inferiore al valore specificato non verranno incluse nel nuovo insieme di regole.

Creazione diretta di un modello di struttura ad albero

In alternativa all'utilizzo del builder della struttura ad albero interattiva, è possibile creare un modello di struttura ad albero delle decisioni direttamente dal nodo durante l'esecuzione del flusso. Questo è coerente con la maggior parte degli altri nodi di creazione dei modelli. Per i modelli di struttura ad albero C5.0, che non sono supportati dal builder della struttura ad albero interattiva, questo è l'unico metodo utilizzabile.

1. Creare un flusso ed aggiungere uno dei nodi della struttura ad albero delle decisioni — C&R Tree, CHAID, QUEST oppure C5.0.
2. Per C&R Tree, QUEST o CHAID, nel pannello Obiettivo della scheda Opzioni di creazione, selezionare uno degli obiettivi principali. Se si sceglie Crea una singola struttura ad albero, assicurarsi che la modalità impostata sia **Genera modello**.

   Per le strutture ad albero C5.0, nella scheda Modello impostare **Tipo di output** su **Struttura ad albero delle decisioni**.
3. Selezionare i campi obiettivo e predittore e definire le opzioni aggiuntive del modello desiderate. Per istruzioni specifiche, consultare la documentazione di ogni nodo di creazione della struttura ad albero.
4. Eseguire il flusso per generare il modello.

Commenti

- Quando si generano strutture ad albero con questo metodo, le direttive di espansione della struttura ad albero vengono ignorate.
- I modelli generati da entrambi i metodi di creazione delle strutture ad albero delle decisioni (interattivo o diretto) sono simili. La scelta del metodo dipende essenzialmente dal tipo di controllo che si desidera esercitare durante la procedura.

Nodi della struttura ad albero delle decisioni

I nodi della struttura ad albero delle decisioni di IBM SPSS Modeler consentono di accedere agli algoritmi di creazione della struttura ad albero descritti in precedenza:

- C&R Tree
- QUEST
- CHAID
- C5.0

Per ulteriori informazioni, consultare l'argomento “Modelli di struttura ad albero delle decisioni” a pagina 81.

Gli algoritmi sono simili poiché creano tutti una struttura ad albero delle decisioni suddividendo in modo ricorsivo i dati in sottogruppi sempre più piccoli. Esistono tuttavia alcune importanti differenze.

Campi di input. I campi di input (predittori) possono essere dei tipi riportati di seguito (livelli di misurazione): continui, relativi alla categoria, indicatori, nominali oppure ordinali.

Campi obiettivo. È possibile specificare un solo campo obiettivo. Per C&R Tree e CHAID, l'obiettivo può essere continuo, relativo alla categoria, indicatore, nominale oppure ordinale. Per QUEST, l'obiettivo può essere categoriale, flag o nominale. Per C5.0, l'obiettivo può essere flag, nominale o ordinale.

Tipo di suddivisione. C&R Tree e QUEST supportano solo suddivisioni binarie (ciascun nodo della struttura ad albero può essere suddiviso in non più di due rami). Al contrario, CHAID e C5.0 supportano la suddivisione in più di due rami per volta.

Metodo utilizzato per la suddivisione. Gli algoritmi sono diversi in base ai criteri utilizzati per decidere il tipo di suddivisione. Quando C&R Tree prevede un output relativo alla categoria, viene utilizzata una misura di dispersione (per impostazione predefinita, il coefficiente Gini, che può comunque essere modificato). Per i target continui, si utilizza il metodo di deviazione quadrata minima. CHAID utilizza un test chi-quadrato, QUEST un test chi-quadrato per predittori categoriali e l'analisi della varianza per gli input continui. Per C5.0 viene utilizzata una misura teorica delle informazioni, il rapporto guadagno informazioni.

Gestione dei valori mancanti. Tutti gli algoritmi consentono valori mancanti per i campi predittori, benché utilizzino metodi diversi per gestirli. C&R Tree e QUEST utilizzano i campi di previsione sostitutivi, quando necessario, per anticipare un record con valori mancanti attraverso la struttura ad albero durante l'addestramento. CHAID rende i valori mancanti una categoria separata e ne consente l'utilizzo nella creazione della struttura ad albero. C5.0 utilizza un metodo di frazionamento, che passa una porzione di record in ogni ramo della struttura ad albero partendo da un nodo in cui la suddivisione è basata su un campo con un valore mancante.

Taglio. C&R Tree, QUEST e C5.0 consentono di ingrandire completamente la struttura ad albero e di tagliarla rimuovendo le suddivisioni di livello più basso che non contribuiscono in modo significativo all'accuratezza della struttura ad albero. Tuttavia, tutti gli algoritmi della struttura ad albero delle decisioni consentono di controllare le dimensioni minime dei sottogruppi, in modo da evitare la creazione di rami con pochi record di dati.

Creazione di strutture ad albero interattive. C&R Tree, QUEST e CHAID dispongono di un'opzione per l'avvio di una sessione interattiva. Ciò consente di creare la struttura ad albero un livello alla volta, modificare le suddivisioni e tagliare la struttura ad albero prima di creare il modello. C5.0 non dispone di un'opzione interattiva.

Probabilità a priori. C&R Tree e QUEST supportano la specifica delle probabilità a priori per le categorie durante la previsione di un campo obiettivo relativo alla categoria. Le probabilità a priori sono stime della frequenza relativa globale di ciascuna categoria obiettivo nella popolazione dalla quale sono estratti i dati di addestramento. In altre parole, sono le stime di probabilità che verrebbero fatte per ciascun possibile valore obiettivo prima di sapere qualcosa sui valori predittori. CHAID e C5.0 non supportano la specifica delle probabilità a priori.

Insiemi di regole. Per i modelli con campi obiettivo categoriali, i nodi della struttura ad albero delle decisioni consentono di creare il modello sotto forma di insieme di regole, che in alcuni casi risulta più semplice da interpretare rispetto a una struttura ad albero delle decisioni complessa. Per C&R Tree, QUEST e CHAID è possibile generare un insieme di regole da una sessione interattiva; per C5.0 è possibile specificare questa opzione sul nodo di modellazione. Inoltre, tutti i modelli di struttura ad albero delle decisioni consentono di generare un insieme di regole dal nugget del modello. Per ulteriori informazioni, consultare l'argomento “Generazione di un Insieme di regole da una struttura ad albero delle decisioni” a pagina 93.

Nodo C&R Tree

Il nodo Classification and Regression (C&R) Tree è un metodo di previsione e classificazione basato sulla struttura ad albero. Questo metodo, analogamente a C5.0, utilizza l'esecuzione ricorsiva di partizioni per suddividere i record in segmenti con valori di campo di output simili. Il nodo C&R Tree esamina i campi di input per individuare la migliore suddivisione, misurata in base alla riduzione in un indice di impurità che risulta dalla suddivisione. La suddivisione definisce due sottogruppi, ognuno dei quali viene successivamente suddiviso in altri due sottogruppi, e così via, fino all'attivazione di un criterio di arresto. Tutte le suddivisioni sono binarie (solo due sottogruppi).

Taglio

Le strutture ad albero C&R consento innanzitutto di ingrandire la struttura ad albero e quindi di eseguire dei tagli in base ad un algoritmo di complessità del costo che regola la stima del rischio in base al numero di nodi terminale. Questo metodo, che consente la crescita della struttura ad albero prima dell'eliminazione in base a criteri più complessi, può avere come risultato strutture ad albero di dimensioni inferiori con migliori proprietà di convalida incrociata. Se si aumenta il numero di nodi terminali, è in genere possibile ridurre il rischio per i dati (di addestramento) correnti, ma il rischio effettivo può risultare maggiore se la generalizzazione eseguita dal modello è applicata a dati non visibili. Si consideri, per esempio, il caso estremo in cui è presente un nodo terminale distinto per ogni record del set di addestramento. La stima del rischio sarà uguale a 0% poiché ogni record è contenuto in un singolo nodo, ma il rischio di errata classificazione per i dati (di test) non visibili sarà quasi certamente maggiore di 0. La misura di complessità del costo rappresenta in questo caso un tentativo di compensazione.

Esempio. Un'emittente televisiva via cavo ha commissionato un'indagine di marketing per determinare quali clienti acquisterebbero un abbonamento a un servizio di notizie interattivo via cavo. Utilizzando i dati dell'indagine è possibile creare un flusso in cui il campo obiettivo è la propensione all'acquisto dell'abbonamento e i campi predittore comprendono età, sesso, livello di istruzione, categoria di reddito, ore passate a guardare la televisione ogni giorno e numero di figli. Applicando un nodo C&R Tree al flusso, sarà possibile prevedere e classificare le risposte in modo da ottenere la percentuale di risposta più alta per la propria campagna.

Requisiti. Per addestrare un modello C&R Tree sono necessari uno o più campi *Input* ed esattamente un campo *Obiettivo*. I campi obiettivo e di input possono essere continui (intervallo numerico) o categoriali. I campi impostati su *Entrambe* o *Nessuna* verranno ignorati. I tipi dei campi utilizzati nel modello devono essere completamente istanziati e i campi ordinali (insieme ordinato) utilizzati nel modello devono includere una classe di archiviazione numerica e non di tipo stringa. Se è necessario convertirli, è possibile utilizzare il nodo Ricodifica.

Efficacia. I modelli C&R Tree sono molto solidi in presenza di problemi come, ad esempio, mancanza di dati e numero elevato di campi. In genere, per la stima di tali modelli non sono necessari tempi di addestramento lunghi. Inoltre, i modelli C&R Tree sono più semplici da comprendere rispetto ad altri tipi di modelli - le regole derivate dal modello sono di interpretazione molto diretta. A differenza del modello C5.0, il modello C&R Tree può contenere sia campi di output continui che categoriali.

Nodo CHAID

CHAID, acronimo di Chi-squared Automatic Interaction Detection, è un metodo di classificazione per la creazione di strutture ad albero delle decisioni basato sull'utilizzo di statistiche chi-quadrato per identificare suddivisioni ottimali.

CHAID analizza innanzitutto le tavole di contingenza tra ognuno dei campi di input e il risultato e quindi verifica la significatività mediante un test di indipendenza chi-quadrato. Se più relazioni sono statisticamente significative, CHAID selezionerà il campo di input più significativo (con il valore *p* più piccolo). Se un input ha più di due categorie, tali categorie vengono confrontate e quelle che non presentano alcuna differenza nei risultati vengono unite insieme. A tale scopo, vengono unite in successione le coppie di categorie che presentano la differenza meno significativa. Il processo di unione

delle categorie si interrompe quando la differenza tra tutte le categorie rimanenti è uguale a quella specificata dal test. Per i campi di input nominali è possibile unire qualsiasi categoria, mentre per gli insiemi ordinali è possibile unire solo le categorie contigue.

Un CHAID completo è una modificazione di CHAID che esegue operazioni avanzate per l'analisi di tutte le suddivisioni possibili per ogni predittore, ma richiede tempi di elaborazione maggiori.

Requisiti. I campi obiettivo e di input possono essere continui o categoriali. I nodi possono essere suddivisi in due o più sottogruppi a ogni livello. I campi ordinali utilizzati nel modello devono includere una classe di archiviazione numerica e non di tipo stringa. Se è necessario convertirli, è possibile utilizzare il nodo Ricodifica.

Efficacia. A differenza dei nodi C&R Tree e QUEST, CHAID può generare strutture ad albero non binarie; ciò significa che alcune suddivisioni dispongono di più di due rami. Tende pertanto a creare strutture ad albero di dimensioni maggiori rispetto ai metodi di crescita binari. CHAID è applicabile a tutti i tipi di input e accetta sia i pesi di caso sia le variabili di frequenza.

Nodo QUEST

QUEST —o Quick, Unbiased, Efficient Statistical Tree— è un metodo di classificazione binario per la creazione di strutture ad albero delle decisioni. Tale metodo è stato sviluppato principalmente per ridurre il tempo di elaborazione necessario per le analisi C&R Tree di grandi dimensioni con molte variabili o molti casi. Un altro obiettivo del metodo QUEST è quello di ridurre la tendenza dei metodi per le strutture ad albero di classificazione a favorire gli input che consentono un numero maggiore di suddivisioni, ovvero i campi di input continui (intervallo numerico) o quelli con un numero elevato di categorie.

- QUEST utilizza una sequenza di regole basate su test di significatività per valutare i campi di input in un nodo. Per la selezione, può essere necessario eseguire almeno un test su ogni input in un nodo. A differenza di C&R Tree, non vengono esaminate tutte le suddivisioni e a differenza di C&R Tree e CHAID, le combinazioni di categorie non vengono verificate durante la valutazione di un campo di input per la selezione. Ciò consente di aumentare la velocità dell'analisi.
- Per determinare le suddivisioni, sui gruppi contenenti le categorie obiettivo viene eseguita l'analisi discriminante quadratica mediante l'input selezionato. Questo metodo consente anch'esso di migliorare la velocità della ricerca completa (C&R Tree) per determinare la suddivisione ottimale.

Requisiti. I campi di input possono essere continui (intervalli numerici), ma il campo obiettivo deve essere categoriale. Tutte le suddivisioni sono binarie. e non è possibile utilizzare i campi peso. I campi ordinali (insieme ordinato) utilizzati nel modello devono includere una classe di archiviazione numerica e non di tipo stringa. Se è necessario convertirli, è possibile utilizzare il nodo Ricodifica.

Efficacia. Come CHAID, ma a differenza di C&R Tree, QUEST utilizza test statistici per decidere se un campo di input viene utilizzato o meno. Separa inoltre le problematiche relative alla selezione dell'input e alla suddivisione, applicando criteri diversi a ognuna di esse. Con CHAID invece, il risultato del test statistico che determina la selezione della variabile crea anche la suddivisione. Allo stesso modo, C&R Tree utilizza la misura di modifica delle impurità per selezionare il campo di input e per determinare la suddivisione.

Opzioni dei campi dei nodi della struttura ad albero delle decisioni

Nella scheda Campi è possibile indicare se si desidera utilizzare le impostazioni dei ruoli dei campi già definite nei nodi a monte oppure creare manualmente le assegnazioni dei campi.

Utilizza ruoli predefiniti. Questa opzione utilizza le impostazioni dei ruoli (obiettivi, predittori e così via) ottenute da un nodo Tipo a monte o dalla scheda Tipi di un nodo origine a monte.

Utilizza assegnazioni campi personalizzate. Scegliere questa opzione per assegnare manualmente obiettivi, predittori e altri ruoli in questa schermata.

Campi. Utilizzare i pulsanti con le frecce per assegnare manualmente le voci dell'elenco ai vari campi dei ruoli situati a destra sullo schermo. Le icone indicano i livelli di misurazione validi per ogni campo di ruolo.

Fare clic sul pulsante **Tutto** per selezionare tutti i campi dell'elenco o fare clic su un singolo pulsante di livello di misurazione per selezionare tutti i campi con tale livello.

Obiettivo. Scegliere un campo come obiettivo per la previsione.

Predittori (input). Scegliere uno o più campi come input per la previsione.

Peso analisi. (Solo CHAID e C&RT) Per utilizzare un campo come peso del caso, specificare il campo in questo punto. I pesi dei casi si utilizzano per tenere conto delle differenze nella varianza tra i vari livelli del campo di output. Per ulteriori informazioni, consultare l'argomento "Utilizzo dei campi frequenza e peso." a pagina 33.

Opzioni di creazione dei nodi della struttura ad albero delle decisioni

La scheda Opzioni di creazione contiene tutte le opzioni che consentono di creare il modello. È possibile, ovviamente, fare clic sul pulsante **Esegui** per creare un modello con tutte le opzioni di default, ma normalmente si rende necessario personalizzare il modello in base alle proprie esigenze.

È possibile indicare qui se creare un nuovo modello o aggiornarne uno già esistente. È anche possibile impostare l'obiettivo principale del nodo: creare un modello standard, creare un nodo con maggiore stabilità o accuratezza o creare un nodo per l'utilizzo con dataset di grandi dimensioni.

Come si desidera procedere?

Crea nuovo modello. Crea un modello completamente nuovo ogni volta che si esegue un flusso contenente questa modalità di creazione del modello (default).

Addestramento continuo modello esistente. Per default, a ogni esecuzione di un nodo Modelli viene creato un modello completamente nuovo. Se si seleziona questa opzione, l'addestramento continuerà con l'ultimo modello creato correttamente dal nodo. In questo modo sarà possibile aggiornare un modello esistente senza dover accedere ai dati originali, aumentando così significativamente le prestazioni poiché nel flusso verranno utilizzati *solo* i record nuovi o aggiornati. I dettagli del modello precedente vengono archiviati con il nodo Modelli, consentendo di utilizzare questa opzione anche se il nugget del modello precedente non è più disponibile nel flusso o nella palette Modelli.

Nota: questa opzione viene attivata solo se si seleziona **Crea un modello per dataset di grandi dimensioni** come obiettivo.

Qual è l'obiettivo principale?

- **Crea una singola struttura ad albero.** Consente di creare un solo modello di struttura ad albero delle decisioni standard. I modelli standard in genere sono più facili da interpretare e consentono un più rapido calcolo del punteggio rispetto ai modelli creati utilizzando le altre opzioni relative all'obiettivo.

  Modalità. Specifica il metodo utilizzato per la creazione del modello. **Genera modello** crea automaticamente un modello quando viene eseguito il flusso. **Avvia sessione interattiva** apre il builder della struttura ad albero, che consente di creare la struttura ad albero un livello alla volta, modificare le suddivisioni ed eseguire le riduzioni desiderate prima di creare il nugget del modello.

  Utilizza direttive della struttura ad albero. Selezionare questa opzione per specificare le direttive da applicare durante la generazione di una struttura ad albero interattiva dal nodo. Per esempio, è

possibile specificare le suddivisioni di primo e secondo livello e applicarle direttamente all'avvio del builder della struttura ad albero. È inoltre possibile salvare le direttive di una sessione di creazione di una struttura ad albero interattiva, allo scopo di ricreare la struttura ad albero in un momento successivo. Per ulteriori informazioni, consultare l'argomento "Aggiornamento delle direttive della struttura ad albero" a pagina 92.

- **Migliorare l'accuratezza del modello (boosting).** Scegliere questa opzione se si desidera utilizzare un metodo speciale, noto come **boosting**, per migliorare il livello di precisione del modello. Il boosting consente di generare modelli multipli in sequenza. Il primo modello viene generato nel modo solito. Viene quindi generato un secondo modello che si concentra sui record classificati erroneamente dal primo modello. Quindi viene generato un terzo modello che si occupa degli errori del secondo modello, e così via. Infine, i casi vengono classificati applicando loro l'intero insieme di modelli, utilizzando una procedura di votazione ponderata per combinare le previsioni separate in un'unica previsione complessiva. Il boosting può migliorare notevolmente la precisione di un modello di struttura ad albero delle decisioni, ma richiede anche un addestramento più lungo.
- **Migliorare la stabilità del modello (bagging).** Scegliere questa opzione se si desidera utilizzare un metodo speciale, noto come **bagging** (bootstrap aggregating), per migliorare la stabilità del modello ed evitare il sovradattamento. Questa opzione consente di creare più modelli e di combinarli in modo tale da ottenere previsioni più affidabili. La creazione dei modelli così ottenuti e il relativo calcolo del punteggio possono richiedere più tempo rispetto ai modelli standard.
- **Crea un modello per dataset di grandi dimensioni.** Scegliere questa opzione se si utilizzano insiemi di dati troppo grandi per la creazione di un modello mediante le altre opzioni relative all'obiettivo. Questa opzione suddivide i dati in blocchi più piccoli e crea un modello su ciascun blocco. I modelli più precisi vengono quindi selezionati automaticamente e combinati in un unico nugget del modello. È possibile eseguire un aggiornamento incrementale del modello se si seleziona l'opzione **Addestramento continuo modello esistente** in questa schermata. *Nota*: l'opzione relativa i dataset di grandi dimensioni richiede una connessione a IBM SPSS Modeler Server.

Nodi della struttura ad albero delle decisioni - opzioni Di base

In questa sezione vengono specificate le opzioni di base relative alle modalità di creazione della struttura ad albero delle decisioni.

Algoritmo di espansione della struttura ad albero. (solo CHAID) Scegliere il tipo di algoritmo **CHAID** da utilizzare. **CHAID completo** è una modificazione di CHAID che esegue operazioni avanzate per l'analisi di tutte le suddivisioni possibili per ogni predittore, ma richiede tempi di elaborazione maggiori.

Profondità massima della struttura ad albero. Specificare il numero massimo di livelli sottostanti al nodo root, ovvero il numero di suddivisioni ricorsive del campione che verranno eseguite. L'impostazione di default è 5. Scegliere **Personalizzato** e immettere un valore per specificare un altro numero di livelli.

Taglio (solo C&RT e QUEST)

Taglia struttura ad albero per evitare sovradattamento. Tagliare una struttura ad albero significa rimuovere le suddivisioni di livello inferiore che non danno un contributo significativo alla precisione della struttura ad albero. Una struttura ad albero tagliata può risultare più semplice e più facile da interpretare e in alcuni casi migliorare la generalizzazione. Se si desidera la struttura ad albero intera, lasciare questa opzione deselezionata.

- **Differenza massima di rischio (in errori standard).** Consente di specificare una regola di taglio più liberale. La regola dell'errore standard consente all'algoritmo di selezionare la struttura ad albero più semplice la cui stima del rischio è vicina (ma possibilmente maggiore) a quella della sottostruttura ad albero con il rischio più basso. Il valore indica la dimensione della differenza ammessa nella stima del rischio tra la struttura ad albero tagliata e quella con il rischio minimo in termini di stima del rischio. Ad esempio, se si specifica 2, è possibile selezionare una struttura ad albero la cui stima del rischio è (2 × errore standard) maggiore di quella della struttura ad albero completa.

Numero massimo surrogati. I surrogati sono un metodo per risolvere il problema dei valori mancanti. Per ogni suddivisione nella struttura ad albero, l'algoritmo identifica i campi di input più simili al campo di suddivisione selezionato. Questi campi sono i **surrogati** per quella suddivisione. Quando è necessario classificare un record ma il record ha un valore mancante per un campo suddiviso, per eseguire la suddivisione è possibile utilizzare il suo valore su un campo surrogato. Se si aumenta questa impostazione, si ottiene una maggiore flessibilità nella gestione dei valori mancanti ma un valore più elevato potrebbe incrementare l'utilizzo della memoria e il tempo di addestramento.

Nodi della struttura ad albero delle decisioni - Regole di arresto

Queste opzioni controllano il modo in cui la struttura ad albero viene creata. Le regole di interruzione determinano quando interrompere la suddivisione di rami specifici della struttura ad albero. Impostare la dimensione minima dei rami per evitare suddivisioni che creerebbero sottogruppi molto piccoli. **Numero min. di record per ramo padre** impedisce la suddivisione se il numero di record nel nodo da suddividere (il **padre**) è inferiore al valore specificato. **Numero min. record per ramo figlio** impedirà la suddivisione se il numero di record in un qualsiasi ramo creato dalla suddivisione (il **figlio**) è inferiore al valore specificato.

- **Utilizza percentuale.** Permette di specificare la dimensione in termini di percentuale dei dati di addestramento globali.
- **Utilizza valore assoluto.** Permette di specificare la dimensione come numero assoluto di record.

Nodi della struttura ad albero delle decisioni - Insiemi

Queste impostazioni determinano il comportamento dei classificatori binari che si verificano quando negli obiettivi sono richiesti boosting, bagging o insiemi di dati di grandi dimensioni. Le opzioni non applicate all'obiettivo selezionato vengono ignorate.

Bagging e insiemi di dati di grandi dimensioni. Quando si calcola il punteggio di un insieme, questo tipo di regola consente di combinare i valori previsti provenienti dai modelli di base per calcolare il valore del punteggio dell'insieme.

- **Regola di combinazione di default per gli obiettivi categoriali.** I valori previsti dell'insieme per gli obiettivi categoriali possono essere combinati utilizzando il confronto, la probabilità massima o la probabilità media più elevata. La **votazione** consente di selezionare la categoria che presenta più spesso la probabilità più elevata nei modelli di base. La **probabilità più elevata** consente di selezionare la categoria che raggiunge la singola probabilità più elevata in tutti i modelli di base. La **probabilità media più elevata** consente di selezionare la categoria con il massimo valore quando viene calcolata la media delle probabilità della categoria in tutti i modelli di base.
- **Regola di combinazione di default per target continui.** I valori previsti degli insiemi per i target continui possono essere combinati utilizzando la media o la mediana dei valori previsti ricavati dai modelli di base.

Si noti che quando l'obiettivo è di ottimizzare la precisione del modello, le selezioni delle regole di combinazione vengono ignorate. Nel boosting viene sempre utilizzato un voto di maggiore ponderazione per il calcolo del punteggio degli obiettivi categoriali e una mediana pesata per il calcolo del punteggio dei target continui.

Boosting e bagging. Specificare il numero dei modelli di base da creare quando l'obiettivo è di ottimizzare la precisione o la stabilità del modello. Per il bagging, si tratta del numero di campioni di bootstrap. Questo valore deve essere un numero intero positivo.

Nodi C&R Tree e QUEST - Costi e probabilità a priori

Costi classificazione errata

In alcuni contesti, certi tipi di errori rappresentano un costo maggiore rispetto ad altri. Potrebbe essere più costoso, per esempio, classificare come a basso rischio chi richiede un fido ad alto rischio (un tipo di

errore) di quanto non lo sia classificare come ad alto rischio chi chiede un fido a basso rischio (un tipo di errore diverso). I costi di errata classificazione permettono di specificare l'importanza relativa dei diversi tipi di errori di previsione.

Essenzialmente i costi di errata classificazione sono pesi applicati a risultati specifici. Tali pesi vengono fattorizzati nel modello e possono realmente modificare la previsione (come modo per proteggersi da errori che gravano sui costi).

Ad eccezione dei modelli C5.0, i costi di errata classificazione non sono applicati quando si calcola il punteggio di un modello e non vengono presi in considerazione quando si classificano o confrontano modelli con il nodo Classificatore automatico, un nodo Analisi o un grafico di valutazione. È possibile che un modello che include costi non generi un numero inferiore di errori rispetto a uno che non li include e che non si classifichi meglio in termini di precisione complessiva, ma è probabile che fornisca prestazioni migliori in termini pratici poiché include una distorsione incorporata a favore degli errori *meno costosi*.

La matrice dei costi mostra il costo per ciascuna possibile combinazione di categoria prevista e categoria effettiva. Per default, tutti i costi di errata classificazione sono impostati su 1.0. Per immettere valori di costi personalizzati, selezionare **Utilizza costi di errata classificazione** e immettere i valori personalizzati nella matrice dei costi.

Per cambiare un costo di errata classificazione, selezionare la cella corrispondente alla combinazione desiderata di valori previsti e valori effettivi, eliminare il contenuto esistente della cella e immettere il costo desiderato. I costi non sono simmetrici automaticamente. Se si imposta, per esempio, il costo dell'errata classificazione di *A* come *B* su 2.0 e non viene esplicitamente cambiato il costo dell'errata classificazione di *B* come *A*, esso manterrà il valore di default 1.0.

Probabilità a priori

Tali opzioni consentono di specificare probabilità a priori per categorie durante la previsione di un campo obiettivo categoriale. Le **probabilità a priori** sono stime della frequenza relativa globale di ciascuna categoria obiettivo nella popolazione dalla quale sono estratti i dati di addestramento. In altre parole, sono le stime di probabilità che verrebbero fatte per ciascun possibile valore obiettivo *prima* di sapere qualcosa sui valori predittori. Esistono tre metodi per impostare le probabilità a priori.

- **In base ai dati di addestramento.** Questa è l'opzione di default. Le probabilità a priori si basano sulle frequenze relative delle categorie nei dati di addestramento.
- **Uguali per tutte le classi.** Le probabilità a priori per tutte le categorie sono definite come $1/k$, dove k è il numero di categorie obiettivo.
- **Personalizzato.** È possibile specificare probabilità a priori personalizzate. L'impostazione dei valori iniziali per le probabilità a priori è uguale per tutte le classi. È possibile impostare le probabilità per le singole categorie su valori definiti dall'utente. Per regolare la probabilità di una categoria specifica, selezionare la cella della probabilità nella tabella corrispondente alla categoria desiderata, eliminare il contenuto della cella e immettere il valore desiderato.

Le probabilità a priori per tutte le categorie devono sommare 1.0 (il **vincolo di probabilità**). Se non assommano a 1.0 viene visualizzato un avviso e viene offerta la possibilità di normalizzare automaticamente i valori. Questa modifica automatica preserva le proporzioni tra le varie categorie e al contempo applica il vincolo di probabilità. Tale modifica può essere eseguita in qualsiasi momento facendo clic sul pulsante **Normalizza**. Per riportare la tabella su valori uguali per tutte le categorie, fare clic sul pulsante **Equalizza**.

Adegua le probabilità a priori utilizzando i costi di errata classificazione. Questa opzione consente di adeguare le distribuzioni di probabilità a priori, in base ai costi di errata classificazione (specificati nella scheda Costi). Questo adeguamento consente di incorporare direttamente nel processo di espansione della struttura ad albero le informazioni sul costo per le strutture ad albero che utilizzano la misura di

impurità Twoing. Se l'opzione non è selezionata, le informazioni sui costi vengono utilizzate solo per la classificazione dei record e per il calcolo delle stime del rischio per le strutture ad albero, in base alla misura Twoing.

Nodo CHAID - Costi

In alcuni contesti, certi tipi di errori rappresentano un costo maggiore rispetto ad altri. Potrebbe essere più costoso, per esempio, classificare come a basso rischio chi richiede un fido ad alto rischio (un tipo di errore) di quanto non lo sia classificare come ad alto rischio chi chiede un fido a basso rischio (un tipo di errore diverso). I costi di errata classificazione permettono di specificare l'importanza relativa dei diversi tipi di errori di previsione.

Essenzialmente i costi di errata classificazione sono pesi applicati a risultati specifici. Tali pesi vengono fattorizzati nel modello e possono realmente modificare la previsione (come modo per proteggersi da errori che gravano sui costi).

Ad eccezione dei modelli C5.0, i costi di errata classificazione non sono applicati quando si calcola il punteggio di un modello e non vengono presi in considerazione quando si classificano o confrontano modelli con il nodo Classificatore automatico, un nodo Analisi o un grafico di valutazione. È possibile che un modello che include costi non generi un numero inferiore di errori rispetto a uno che non li include e che non si classifichi meglio in termini di precisione complessiva, ma è probabile che fornisca prestazioni migliori in termini pratici poiché include una distorsione incorporata a favore degli errori *meno costosi*.

La matrice dei costi mostra il costo per ciascuna possibile combinazione di categoria prevista e categoria effettiva. Per default, tutti i costi di errata classificazione sono impostati su 1.0. Per immettere valori di costi personalizzati, selezionare **Utilizza costi di errata classificazione** e immettere i valori personalizzati nella matrice dei costi.

Per cambiare un costo di errata classificazione, selezionare la cella corrispondente alla combinazione desiderata di valori previsti e valori effettivi, eliminare il contenuto esistente della cella e immettere il costo desiderato. I costi non sono simmetrici automaticamente. Se si imposta, per esempio, il costo dell'errata classificazione di *A* come *B* su 2.0 e non viene esplicitamente cambiato il costo dell'errata classificazione di *B* come *A*, esso manterrà il valore di default 1.0.

Nodo C&R Tree - Opzioni avanzate

Le opzioni avanzate consentono di definire con precisione il processo di creazione delle strutture ad albero.

Modifica minima in impurità. Specificare la modifica minima in impurità per creare una nuova suddivisione nella struttura ad albero. **Impurità** si riferisce al grado di presenza, nei sottogruppi definiti dalla struttura ad albero, di un ampio intervallo di valori di campi di output all'interno di ciascun gruppo. Per i target di categoria, un nodo viene considerato “puro” se il 100% dei casi nel nodo rientrano in una specifica categoria del campo obiettivo. La creazione della struttura ad albero è finalizzata alla creazione di sottogruppi con valori di output simili, ovvero alla riduzione al minimo dell'impurità all'interno di ogni nodo. Se la suddivisione migliore per un ramo riduce l'impurità della struttura ad albero di un valore inferiore a quello specificato, la suddivisione non verrà eseguita.

Misura di impurità per obiettivi categoriali. Per i campi obiettivo categoriali, specificare il metodo utilizzato per misurare l'impurità della struttura ad albero. Per i target continui, questa opzione viene ignorata e viene sempre utilizzata la misura di impurità **deviazione quadrata minima**.

- **Gini** è una misura di impurità generale basata sulle probabilità dell'appartenenza a una categoria per il ramo.
- **Twoing** è una misura di impurità che enfatizza la suddivisione binaria e che è più probabile porti a rami di dimensioni approssimativamente uguali in conseguenza di una suddivisione.

- **Ordinato** aggiunge il vincolo che è possibile raggruppare solo le classi obiettivo contigue, poiché è applicabile unicamente con gli obiettivi ordinali. Se l'opzione viene selezionata per un obiettivo nominale, per default viene utilizzata la misura Twoing standard.

Insieme di prevenzione del sovradattamento. L'algoritmo consente di separare internamente i record in un insieme di creazione del modello e un insieme di prevenzione del sovradattamento, che è un insieme indipendente di record di dati utilizzato per tenere traccia degli errori durante l'addestramento per impedire al metodo la modellazione della variazione casuale nei dati. Specificare una percentuale di record. Il valore predefinito è 30.

Replica risultati. L'impostazione di un seed random consente di replicare le analisi. Specificare un intero o fare clic su **Genera** per creare un intero pseudocasuale compreso tra 1 e 2147483647 incluso.

Nodo QUEST - Opzioni avanzate

Le opzioni avanzate consentono di definire con precisione il processo di creazione delle strutture ad albero.

Livello di significatività per suddivisione. Specifica il livello di significatività (alfa) per la suddivisione dei nodi. Il valore deve essere compreso tra 0 e 1. Valori inferiori tendono a produrre strutture ad albero con un numero minore di nodi.

Insieme di prevenzione del sovradattamento. L'algoritmo consente di separare internamente i record in un insieme di creazione del modello e un insieme di prevenzione del sovradattamento, che è un insieme indipendente di record di dati utilizzato per tenere traccia degli errori durante l'addestramento per impedire al metodo la modellazione della variazione casuale nei dati. Specificare una percentuale di record. Il valore predefinito è 30.

Replica risultati. L'impostazione di un seed random consente di replicare le analisi. Specificare un intero o fare clic su **Genera** per creare un intero pseudocasuale compreso tra 1 e 2147483647 incluso.

Nodo CHAID - Opzioni avanzate

Le opzioni avanzate consentono di definire con precisione il processo di creazione delle strutture ad albero.

Livello di significatività per suddivisione. Specifica il livello di significatività (alfa) per la suddivisione dei nodi. Il valore deve essere compreso tra 0 e 1. Valori inferiori tendono a produrre strutture ad albero con un numero minore di nodi.

Livello di significatività per unione. Specifica il livello di significatività (alfa) per l'unione delle categorie. Il valore deve essere maggiore di 0 e minore o uguale a 1. Per evitare l'unione di categorie, specificare un valore 1. Per i target continui, ciò significa che il numero di categorie per la variabile nella struttura ad albero finale corrisponde al numero di intervalli specificato. Questa opzione non è disponibile per CHAID completo.

Adegua valori di significatività tramite il metodo di Bonferroni. Adegua i valori di significatività durante la verifica delle diverse combinazioni di categorie di un predittore. I valori vengono adeguati in base al numero di test, che è direttamente correlato al numero di categorie e al livello di misurazione di un predittore. Questa è in genere una soluzione ottimale perché consente di controllare la frequenza di errore falso positivo. Se si disattiva questa opzione, la capacità dell'analisi di trovare le differenze effettive verrà aumentata, ma verrà aumentato anche il tasso dei falsi positivi. È consigliabile disattivare questa opzione per i campioni di piccole dimensioni.

Consenti la risuddivisione delle categorie unite all'interno di un nodo. L'algoritmo CHAID tenta di unire le categorie per creare la struttura ad albero più semplice che descrive il modello. Se selezionata, questa opzione consente di suddividere nuovamente le categorie unite, se tale operazione consente di ottenere una soluzione migliore.

Chi-quadrato per obiettivi categoriali. Per gli obiettivi categoriali, è possibile specificare il metodo utilizzato per calcolare la statistica chi-quadrato.

- **Pearson.** Questo metodo fornisce calcoli più rapidi ma deve essere utilizzato con cautela su campioni di piccole dimensioni.
- **Rapporto di verosimiglianza.** Questo metodo è più solido del metodo di Pearson ma richiede tempi di calcolo più lunghi. È il metodo di elezione per campioni di piccole dimensioni e viene sempre utilizzato per i target continui.

Modifica minima nelle frequenze di cella previste. Per la stima delle frequenze di cella (sia per il modello nominale che per il modello ordinale degli effetti di riga), la convergenza sulla stima ottimale utilizzata nel test chi-quadrato per una suddivisione specifica viene eseguita mediante una procedura iterativa (epsilon). Epsilon determina il grado di cambiamento necessario affinché le iterazioni possano continuare. Se il cambiamento dall'ultima iterazione è inferiore al valore specificato, le iterazioni vengono interrotte. Se si verificano problemi e l'algoritmo non esegue la convergenza, è possibile aumentare questo valore o diminuire il numero massimo di iterazioni fino a quando la convergenza non viene eseguita.

Numero massimo di iterazioni per la convergenza. Specifica il numero massimo di iterazioni prima dell'arresto, indipendentemente dal fatto che la convergenza sia stata eseguita o meno.

Replica risultati. L'impostazione di un seed random consente di replicare le analisi. Specificare un intero o fare clic su **Genera** per creare un intero pseudocasuale compreso tra 1 e 2147483647 incluso.

Opzioni del modello per il nodo Struttura ad albero delle decisioni

Nella scheda Opzioni modello è possibile scegliere se specificare un nome per il modello o generare un nome automaticamente. È inoltre possibile scegliere di ottenere informazioni sull'importanza dei predittori nonché sui punteggi delle propensioni grezze e regolate per gli obiettivi flag.

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.

Valutazione del modello

Calcola importanza predittori. Per i modelli che generano una misura appropriata dell'importanza è possibile visualizzare un grafico che indica l'importanza relativa di ogni predittore nella stima del modello. In genere è opportuno concentrare la modellazione sui predittori più importanti e valutare se eliminare o ignorare quelli meno importanti. Notare che per calcolare l'importanza dei predittori per alcuni modelli potrebbe essere necessario più tempo, in particolare quando vengono utilizzati dataset di grandi dimensioni e che, come risultato, la funzione è disattivata per impostazione predefinita per alcuni modelli. L'importanza dei predittori non è disponibile per i modelli di elenco di decisioni. Per ulteriori informazioni, consultare l'argomento “Importanza predittore” a pagina 43.

Punteggi di propensione

I punteggi di propensione possono essere attivati nel nodo Modelli e nella scheda Impostazioni del nugget del modello. Questa funzionalità è disponibile solamente quando l'obiettivo selezionato è un campo flag. Per ulteriori informazioni, consultare l'argomento “Punteggi di propensione” a pagina 35.

Calcola punteggi di propensione grezza. I punteggi di propensione grezza vengono derivati dal modello in base ai soli dati di addestramento. Se il modello prevede il valore *vero* (risposta favorevole), la propensione sarà uguale a P, dove P è la probabilità della previsione. SE il modello prevede il valore falso, la propensione viene calcolata come (1 – P).

- Se si seleziona questa opzione durante la creazione del modello, nel nugget del modello i punteggi di propensione saranno attivati per default. Tuttavia, nel nugget del modello si può sempre decidere di attivare i punteggi di propensione grezza indipendentemente dal fatto che siano stati selezionati nel nodo Modelli.
- Quando si calcola il punteggio del modello, i punteggi di propensione grezza saranno aggiunti in un campo al cui prefisso standard vengono fatte seguire le lettere *RP*. Per esempio, se le previsioni sono in un campo denominato *$R-tasso di abbandono*, il nome del campo del punteggio di propensione sarà *$RRP-tasso di abbandono*.

Calcola punteggi di propensione regolata. Le propensioni grezze si basano unicamente sulle stime fornite dal modello, le quali possono essere sovradattate e determinare di conseguenza delle stime eccessivamente ottimistiche della propensione. Le propensioni regolate cercano di compensare esaminando le prestazioni del modello sulle partizioni di test e di convalida e adeguando le propensioni di conseguenza per fornire una stima più corretta.

- Questa impostazione richiede la presenza di un campo partizione valido nel flusso.
- Al contrario dei punteggi di confidenza grezza, i punteggi di propensione regolata devono essere calcolati in sede di creazione del modello, altrimenti non saranno disponibili quando si calcola il punteggio del nugget del modello.
- Quando si calcola il punteggio del modello, i punteggi di propensione regolata saranno aggiunti in un campo al cui prefisso standard vengono fatte seguire le lettere *AP*. Per esempio, se le previsioni sono in un campo denominato *$R-tasso di abbandono*, il nome del campo del punteggio di propensione sarà *$RAP-tasso di abbandono*. I punteggi di propensione regolata non sono disponibili per i modelli di regressione logistica.
- Quando si calcolano i punteggi di propensione regolata, la partizione di test o di convalida utilizzata per il calcolo non deve essere stata bilanciata. A tal fine, verificare che l'opzione **Bilancia solo dati di addestramento** sia selezionata in tutti i nodi bilanciamento a monte. Inoltre, se a monte è presente un campione complesso, i punteggi di propensione regolata saranno invalidati.
- I punteggi di propensione regolata non sono disponibili per i modelli di struttura ad albero boosted e di insiemi di regole. Per ulteriori informazioni, consultare l'argomento “Modelli C5.0 boosted” a pagina 113.

In base a. Per consentire il calcolo del punteggio di propensione regolata, nel flusso deve essere presente un campo partizione. È possibile specificare se per il calcolo deve essere utilizzata la partizione di test o di convalida. Per ottenere risultati migliori, la partizione di test o di convalida deve contenere un numero di record pari almeno a quelli presenti nella partizione utilizzata per l'addestramento del modello originale.

nodo C5.0

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

Questo nodo utilizza l'algoritmo C5.0 per costruire una **struttura ad albero delle decisioni** o un **insieme di regole**. Un modello C5.0 suddivide il campione in base al campo che fornisce il massimo **guadagno di informazioni**. Ogni sottocampione definito dalla prima suddivisione viene ulteriormente suddiviso, in genere in base a un campo diverso, e il processo viene ripetuto finché non è più possibile suddividere ulteriormente i sottocampioni. Vengono infine riesaminate le suddivisioni di livello più basso e quelle che non contribuiscono in modo significativo al valore del modello vengono rimosse o **tagliate**.

Nota: il nodo C5.0 può prevedere solo un obiettivo categoriale. Quando si analizzano dati con campi categoriali (nominali o ordinali), è più probabile che il nodo raggruppi insieme delle categorie che nelle versioni di C5.0 precedenti alla 11.0.

C5.0 può produrre due tipi di modelli. Una **struttura ad albero delle decisioni** è una descrizione semplice e diretta delle suddivisioni riscontrate dall'algoritmo. Ogni nodo terminale, o "foglia", descrive

un particolare sottoinsieme dei dati di addestramento e ogni caso nei dati di addestramento appartiene a un nodo terminale specifico nella struttura ad albero. In altre parole, è possibile avere una sola previsione per un qualsiasi particolare record di dati che viene presentato a una struttura ad albero delle decisioni.

Al contrario, un **insieme di regole** cerca di fare previsioni per i singoli record. Gli insiemi di regole derivano dalle strutture ad albero delle decisioni e, in un certo qual modo, rappresentano una versione semplificata o "distillata" delle informazioni trovate nella struttura ad albero delle decisioni. Gli insiemi di regole spesso sono in grado di mantenere le informazioni più importanti di un'intera struttura ad albero delle decisioni ma con un modello meno complesso. Dato il modo in cui operano, gli insiemi di regole non hanno le stesse proprietà delle strutture ad albero delle decisioni. La differenza più importante è il fatto che, con un insieme di regole, a un particolare record può applicarsi più di una regola o nessuna. Se si applicano più regole, ognuna di esse riceve un "voto" ponderato in base alla confidenza associata a quella regola e la previsione finale viene decisa combinando i voti ponderati di tutte le regole che si applicano al record interessato. Se non si applica alcuna regola, al record viene assegnata una previsione di default.

Esempio. Un ricercatore medico ha raccolto dati relativi a un gruppo di pazienti, tutti colpiti dalla stessa malattia. Nel corso della terapia, ogni paziente è stato sottoposto a una cura scelta tra cinque. È possibile utilizzare un modello C5.0, insieme ad altri nodi, per individuare il farmaco adatto per un futuro paziente colpito dalla stessa malattia.

Requisiti. L'addestramento di un modello C5.0 richiede un campo categoriale (cioè nominale o ordinale) *Obiettivo* e uno o più campi *Input* di qualsiasi tipo. I campi impostati su *Entrambe* o *Nessuna* verranno ignorati. È necessario che i tipi dei campi utilizzati nel modello siano completamente istanziati. È anche possibile specificare un campo peso.

Efficacia. I modelli C5.0 sono molto stabili in presenza di problemi quali dati mancanti e grandi numeri di campi di input. In genere, per la stima di tali modelli non sono necessari tempi di addestramento lunghi. Inoltre, i modelli C5.0 tendono a essere più facili da capire rispetto ad altri tipi di modello, dato che le regole da essi derivate sono di interpretazione molto diretta. C5.0 offre anche il potente metodo **boosting** per aumentare la precisione della classificazione.

Nota: la velocità di creazione del modello C5.0 può trarre vantaggio dall'abilitazione dell'elaborazione parallela.

Opzioni del modello di nodo C5.0

Nome modello. Specificare il nome del modello da creare.

- **Auto.** Se questa opzione è selezionata, il nome del modello verrà generato automaticamente in base al nome del campo o dei campi obiettivo. Questa è l'opzione di default.
- **Personalizzato.** Selezionare questa opzione per specificare un nome personalizzato per il nugget del modello che verrà creato da questo nodo.

Utilizza dati partizionati. Se è definito un campo partizione, questa opzione garantisce che per la creazione del modello vengano utilizzati dati solo dalla partizione di addestramento.

Crea modelli di suddivisione. Crea un modello separato per ogni valore possibile dei campi di input che sono stati specificati come campi di suddivisione. Per ulteriori informazioni, consultare l'argomento “Creazione di modelli di suddivisione” a pagina 28.

Tipo di output. Specificare se si desidera che il nugget del modello risultante sia una **Struttura ad albero delle decisioni** o un **Insieme di regole**.

Raggruppa valori simbolici. Se questa opzione è selezionata, C5.0 tenterà di combinare i valori simbolici che hanno schemi simili rispetto al campo di output. Se questa opzione non è selezionata, C5.0 creerà un nodo figlio per ciascun valore del campo simbolico utilizzato per suddividere il nodo padre. Per esempio,

se C5.0 esegue la suddivisione in un campo *COLORE* (con valori *ROSSO*, *VERDE* e *BLU*), creerà per default una suddivisione tridirezionale. Se, però, questa opzione è selezionata e i record in cui *COLORE* = *ROSSO* sono simili a quelli in cui *COLORE* = *BLU*, C5.0 creerà una suddivisione bidirezionale, con i *VERDE* in un gruppo e i *BLU* e i *ROSSO* inseriti insieme nell'altro gruppo.

Utilizza aumento. L'algoritmo C5.0 ha un metodo speciale per migliorare il suo tasso di precisione, chiamato **boosting**. Questo metodo consente di generare modelli multipli in sequenza. Il primo modello viene generato nel modo solito. Viene quindi generato un secondo modello che si concentra sui record classificati erroneamente dal primo modello. Quindi viene generato un terzo modello che si occupa degli errori del secondo modello, e così via. Infine, i casi vengono classificati applicando loro l'intero insieme di modelli, utilizzando una procedura di votazione ponderata per combinare le previsioni separate in un'unica previsione complessiva. Il boosting può migliorare notevolmente la precisione di un modello C5.0, ma richiede anche un addestramento più lungo. L'opzione **Numero di prove** consente di controllare il numero di modelli utilizzati per il modello boosted. Questa funzione è basata sulla ricerca di Freund & Schapire, con alcuni miglioramenti proprietari per la gestione dei dati rumorosi.

Convalida incrociata. Se questa opzione è selezionata, C5.0 utilizzerà un insieme di modelli generati sui sottoinsiemi dei dati di addestramento per stimare la precisione di un modello generato sull'intero dataset. Questa funzione è utile se il dataset è troppo piccolo per essere suddiviso nei tradizionali insiemi di addestramento e test. I modelli di convalida incrociata vengono scartati dopo il calcolo della stima della precisione. È possibile specificare il **numero di occorrenze** o il numero di modelli utilizzati per la convalida incrociata. Si noti che nelle versioni precedenti di IBM SPSS Modeler, la generazione del modello e la convalida incrociata sono due operazioni separate. Nella versione corrente, non è richiesta alcuna fase separata per la generazione del modello. La creazione del modello e la convalida incrociata vengono eseguite contemporaneamente.

Modalità. Per l'addestramento **Livello base**, la maggior parte dei parametri C5.0 viene impostata automaticamente. L'addestramento **Livello avanzato** consente un maggior controllo diretto sui parametri di addestramento.

Opzioni della modalità Livello base Preferenza.

Preferenza. Per default, C5.0 cercherà di produrre la struttura ad albero più precisa possibile. In alcuni casi, questo può portare a un sovradattamento, con conseguente decadimento delle prestazioni quando il modello viene applicato a nuovi dati. Selezionare **Generalità** per utilizzare impostazioni dell'algoritmo meno suscettibili di causare questo problema.

Nota: non esiste alcuna garanzia che i modelli generati con l'opzione **Generalità** vengano generalizzati meglio di altri. Quando la generalità è un requisito importante, validare sempre il modello a fronte di un campione estratto per il test.

Rumore previsto (%). Specificare la percentuale prevista di dati errati o rumorosi nel set di addestramento.

Opzioni della modalità avanzata

Gravità di taglio. Determina la portata del taglio che verrà praticato alla struttura ad albero delle decisioni o all'insieme di regole. Aumentare questo valore per ottenere una struttura ad albero più piccola e concisa. Ridurlo per ottenere una struttura ad albero più precisa. Questa impostazione influisce solo sul taglio locale (vedere "Utilizza taglio globale").

Numero minimo record per ramo figlio. La dimensione dei sottogruppi può essere utilizzata per limitare il numero di suddivisioni in un qualsiasi ramo della struttura ad albero. Un ramo della struttura ad albero verrà suddiviso solo se due o più sottorami risultanti contengono almeno questo numero di record del set di addestramento. Il valore di default è 2. Aumentarlo per prevenire l'**overtraining** con dati rumorosi.

Utilizza taglio globale. Le strutture ad albero vengono tagliate in due fasi: viene prima eseguito un taglio locale che esamina le sottostrutture ad albero e comprime i rami per migliorare la precisione del modello. Quindi, la fase di taglio globale prende in considerazione l'intera struttura ad albero e le sottostrutture ad albero deboli vengono compresse. Il taglio globale viene eseguito per default. Per evitare la fase di taglio globale, deselezionare questa opzione.

Attributi Winnow. Se questa opzione è selezionata, prima di iniziare a generare il modello C5.0 esaminerà l'utilità dei predittori. I predittori che vengono considerati irrilevanti vengono esclusi dal processo di generazione del modello. Questa opzione può essere utile per modelli con molti campi predittori e può contribuire a evitare il sovradattamento.

Nota: la velocità di creazione del modello C5.0 può trarre vantaggio dall'abilitazione dell'elaborazione parallela.

Nugget del modello Struttura ad albero delle decisioni

I nugget del modello Struttura ad albero delle decisioni rappresentano la struttura ad albero per la previsione di un determinato campo di output scoperto da uno dei nodi Modelli della struttura ad albero delle decisioni (C&R Tree, CHAID, QUEST, C5.0). È possibile creare modelli di struttura ad albero direttamente dal nodo di creazione della struttura ad albero o indirettamente dal builder della struttura ad albero interattiva. Per ulteriori informazioni, consultare l'argomento "Builder della struttura ad albero interattiva" a pagina 82.

Calcolo del punteggio dei modelli di struttura ad albero

Quando si esegue un flusso contenente un nugget del modello della struttura ad albero, il risultato specifico dipende dal tipo di struttura ad albero.

- Per le strutture ad albero di classificazione (obiettivo categoriale), ai dati vengono aggiunti due nuovi campi contenenti il valore previsto e la confidenza relativa a ogni record. La previsione si basa sulla categoria più frequente per il nodo terminale al quale viene assegnato il record; se la maggioranza dei rispondenti in un determinato nodo è *sì*, la previsione di tutti i record assegnati a quel nodo sarà sì.
- Per le strutture ad albero di regressione, vengono generati solo i valori previsti, mentre le confidenze non vengono assegnate.
- Se lo si desidera, per i modelli CHAID, QUEST e C&R Tree è possibile aggiungere un altro campo che indica l'ID del nodo al quale è assegnato ogni record.

Per il nome dei nuovi campi viene utilizzato il nome del modello con l'aggiunta di prefissi. Per i nodi C&R Tree, CHAID e QUEST, i prefissi sono *$R-* per il campo della previsione, *$RC-* per il campo della confidenza e *$RI-* per il campo dell'identificatore del nodo. Per le strutture ad albero C5.0, i prefissi sono *$C-* per il campo della previsione e *$CC-* per il campo della confidenza. Se sono presenti più nodi modello della struttura ad albero, i nomi dei nuovi campi conterranno, se necessario, dei numeri nel *prefisso*, per consentire di distinguerli tra loro — ad esempio, *$R1-*, *$RC1-* e *$R2-*.

Utilizzo dei nugget del modello di struttura ad albero

È possibile salvare o esportare le informazioni relative al modello in diversi modi.

Nota: molte di queste opzioni sono disponibili anche nella finestra del Builder della struttura ad albero.

Dal builder della struttura ad albero o da un nugget del modello della struttura ad albero, è possibile:

- Generare un nodo Filtro o Seleziona in base alla struttura ad albero corrente. Per ulteriori informazioni, consultare l'argomento "Generazione di nodi Filtro e Seleziona" a pagina 93.

- Generare un nugget del modello Insieme di regole contenente la struttura ad albero come insieme di regole che definiscono i rami finali della struttura ad albero. Per ulteriori informazioni, consultare l'argomento “Generazione di un Insieme di regole da una struttura ad albero delle decisioni” a pagina 93.
- Inoltre, solo per i nugget del modello della struttura ad albero, è possibile esportare il modello in formato PMML. Per ulteriori informazioni, consultare l'argomento “La palette Modelli” a pagina 40. Se il modello include suddivisioni personalizzate, tali informazioni non vengono conservate nel PMML esportato. (La suddivisione viene conservata, ma non il fatto che sia personalizzata anziché scelta dall'algoritmo).
- Generare un grafico in base a una parte selezionata della struttura ad albero corrente. *Nota*: questa operazione è valida per un nugget solo quando è collegato ad altri nodi in un flusso. Per ulteriori informazioni, consultare l'argomento “Generazione di grafici” a pagina 113.
- Unicamente per i modelli C5.0 boosted, è possibile scegliere **Struttura ad albero delle decisioni singola (area)** oppure **Struttura ad albero delle decisioni singola (palette MG)** per creare un nuovo Insieme di regole derivato dalla regola selezionata corrente. Per ulteriori informazioni, consultare l'argomento “Modelli C5.0 boosted” a pagina 113.

Nota: sebbene il nodo Creazione regola sia stato sostituito dal nodo C&R Tree, i nodi della struttura ad albero delle decisioni originariamente creati utilizzando un nodo Creazione regola continueranno a funzionare correttamente.

Nugget del modello struttura ad albero singola

Se si seleziona l'opzione **Crea una singola struttura ad albero** come obiettivo principale nel nodo di modellazione, il nugget del modello risultante conterrà le seguenti schede.

Tabella 7. Schede nel nugget della struttura ad albero singolo

Scheda	Descrizione	Ulteriori informazioni
Modello	Visualizza le regole che definiscono il modello.	Per ulteriori informazioni, consultare l'argomento “Regole dei modelli di struttura ad albero delle decisioni”.
Visualizzatore	Visualizza la struttura del modello.	Per ulteriori informazioni, consultare l'argomento “Visualizzatore del modello Struttura ad albero delle decisioni” a pagina 111.
Riepilogo	Visualizza informazioni sui campi, le impostazioni di creazione e l'elaborazione della stima del modello.	Per ulteriori informazioni, consultare l'argomento “Scheda Riepilogo di un nugget del modello/Informazioni” a pagina 43.
Impostazioni	Consente di specificare le opzioni per le confidenze e per la generazione SQL durante il calcolo del punteggio del modello.	Per ulteriori informazioni, consultare l'argomento “Impostazioni del nugget del modello Struttura ad albero delle decisioni/Insieme di regole” a pagina 112.
Annotazione	Consente di aggiungere annotazioni descrittive, specificare un nome personalizzato, aggiungere un testo di suggerimento e inserire le parole chiave di ricerca per il modello.	

Regole dei modelli di struttura ad albero delle decisioni

Nella scheda Modello di un nugget della struttura ad albero delle decisioni sono visualizzate le regole che definiscono il modello. Se lo si desidera, è inoltre possibile visualizzare un grafico di importanza dei predittori e un terzo riquadro con informazioni su cronologia, frequenze e surrogati.

Nota: se nella scheda Opzioni di creazione (pannello Obiettivo) del nodo CHAID viene selezionata l'opzione **Crea un modello per insiemi di dati di grandi dimensioni**, la scheda Modello visualizza solo i dettagli della regola della struttura ad albero.

Regole della struttura ad albero

Il riquadro di sinistra visualizza un elenco delle condizioni che definiscono il partizionamento dei dati rilevati dall'algoritmo — essenzialmente una serie di regole che è possibile utilizzare per assegnare record individuali ai nodi figlio in base ai valori di predittori differenti.

Le strutture ad albero delle decisioni funzionano eseguendo partizioni dei dati in modo ricorsivo in base ai valori dei campi di input. Le partizioni dei dati vengono denominate **rami**. Il ramo iniziale (a volte denominato **radice**) include tutti i record dei dati. La radice viene suddivisa in sottoinsiemi o **rami figlio**, in base al valore di un determinato campo di input. Ogni ramo figlio può essere ulteriormente suddiviso in sottorami, che a loro volta possono essere ulteriormente suddivisi, e così via. Al livello più basso della struttura ad albero si trovano i rami senza ulteriori suddivisioni. Questi rami vengono denominati **rami finali** (o **foglie**).

Dettagli delle regole della struttura ad albero

Il browser di regole visualizza i valori di input che definiscono ogni partizione o ramo e un riepilogo dei valori dei campi di output relativi ai record nella suddivisione. Per informazioni generali relative all'utilizzo del browser del modello, consultare “Esplorazione dei nugget del modello” a pagina 42.

Per le suddivisioni basate su campi numerici, il ramo viene visualizzato da una linea della forma seguente:

```
fieldname relation value [summary]
```

dove *relation* è una relazione numerica. Per esempio, un ramo definito da valori maggiori di 100 per il campo *revenue* viene visualizzato come:

```
revenue > 100 [summary]
```

Per le suddivisioni basate su campi numerici, il ramo viene visualizzato da una linea della forma seguente:

```
fieldname = value [summary] or fieldname in [values] [summary]
```

dove *values* rappresenta i valori del campo che definiscono il ramo. Per esempio, un ramo che include record dove il valore di *region* può essere *North*, *West* oppure *South* viene rappresentato come:

```
region in ["North" "West" "South"] [summary]
```

Per i rami finali, viene inoltre eseguita una previsione che aggiunge una freccia e il valore previsto alla fine della condizione della regola. Per esempio, in una foglia definita da *revenue* > 100 che prevede un valore *high* per il campo output viene visualizzato come:

```
revenue > 100 [Mode: high] → high
```

Il valore **summary** relativo al ramo viene definito in modo diverso per i campi di output numerici e simbolici. Per le strutture ad albero con campi di output numerici, il riepilogo rappresenta il valore **average** relativo al ramo e l'**effect** del ramo, la differenza tra la media relativa al ramo e la media relativa al ramo padre. Per le strutture ad albero con campi di output simbolici, il riepilogo rappresenta la **modalità**, o il valore più frequente, dei record nel ramo.

Per descrivere un ramo completamente, è necessario includere la condizione che definisce il ramo in aggiunta alle condizioni che definiscono le ulteriori suddivisioni della struttura ad albero. Per esempio, nella struttura ad albero:

```
revenue > 100
  region = "North"
  region in ["South" "East" "West"]
    revenue <= 200
```

il ramo rappresentato dalla seconda linea viene definito dalle condizioni *revenue > 100* e *region = "North"*.

Se si fa clic su **Mostra occorrenze/confidenza** sulla barra degli strumenti, ogni regola visualizzerà anche le informazioni sul numero di record a cui si applica la regola (**Istanze**) e la proporzione dei record per cui la regola è vera (**Confidenza**).

Importanza predittore

Facoltativamente, nella scheda Modello, è possibile visualizzare anche un grafico che indica l'importanza relativa di ogni predittore nella stima del modello. In genere è opportuno concentrare la modellazione sui predittori più importanti e valutare se eliminare o ignorare quelli meno importanti. Notare che tale grafico è disponibile solo se è selezionata l'opzione **Calcola importanza predittore** nella scheda Analizza prima di generare il modello. Per ulteriori informazioni, consultare l'argomento "Importanza predittore" a pagina 43.

Informazioni aggiuntive sul modello

Se si fa clic su **Visualizza un riquadro di informazioni aggiuntive** sulla barra degli strumenti, nella parte inferiore della finestra viene visualizzato un riquadro contenente informazioni dettagliate per la regola selezionata. Il riquadro delle informazioni contiene tre schede.

Cronologia. In questa scheda vengono tracciate le condizioni di suddivisione dal nodo root fino al nodo selezionato. Si tratta di un elenco di condizioni che stabilisce quando un record viene assegnato al nodo selezionato. I record per i quali tutte le condizioni sono vere verranno assegnati a questo nodo.

Frequenze. Nei modelli con campi obiettivo simbolici, questa scheda visualizza, di ogni valore di destinazione possibile, il numero di record con tale valore di destinazione, assegnati al nodo (nei dati di addestramento). Viene inoltre visualizzata la cifra che indica la frequenza, espressa in termini percentuali (con un massimo di tre numeri decimali). Nei modelli con destinazioni numeriche, questa scheda è vuota.

Surrogati. Se possibile, qualsiasi surrogato per il campo della suddivisione principale viene visualizzato per il nodo selezionato. I surrogati sono campi alternativi utilizzati quando manca il valore del predittore principale per un record determinato. Il numero massimo di surrogati consentito per una determinata suddivisione viene specificato nel nodo di creazione della struttura ad albero ma il numero effettivo dipende dai dati di addestramento. In generale, più dati mancano, più surrogati vengono utilizzati. In altri modelli di strutture ad albero delle decisioni, questa scheda è vuota.

Nota: per essere inclusi nel modello, i surrogati devono essere identificati durante la fase di addestramento. Se il campione di addestramento non ha valori mancanti, nessun surrogato verrà identificato e qualsiasi record con valori mancanti trovato durante i test o il calcolo del punteggio rientra automaticamente nel nodo figlio con il numero maggiore di record. Se sono previsti valori mancanti durante i test o il calcolo del punteggio, accertarsi che i valori siano assenti anche nel campione di addestramento. I surrogati non sono disponibili per le strutture ad albero CHAID.

Visualizzatore del modello Struttura ad albero delle decisioni

La scheda Visualizzatore di un nugget del modello di struttura ad albero delle decisioni è simile alla visualizzazione nel builder della struttura ad albero. La differenza principale è data dal fatto che navigando nel nugget del modello non è possibile ingrandire o modificare la struttura ad albero. Nei due componenti sono disponibili altre opzioni simili per la consultazione e la personalizzazione della visualizzazione. Per ulteriori informazioni, consultare l'argomento "Personalizzazione della vista della struttura ad albero" a pagina 85.

Nota: la scheda Visualizzatore non viene visualizzata per i nugget del modello CHAID creati se si seleziona l'opzione **Crea un modello per insiemi di dati di grandi dimensioni** nella scheda Opzioni di creazione - pannello Obiettivo.

Quando si visualizzano le regole di suddivisione nella scheda Visualizzatore, le parentesi quadre indicano che il valore adiacente è incluso nell'intervallo, mentre le parentesi tonde indicano che il valore adiacente è escluso dall'intervallo. Pertanto, l'espressione (23,37] significa da 23 escluso a 37 incluso, cioè da appena sopra a 23 fino a 37. Nella scheda Modello, la stessa condizione sarebbe visualizzata come:

```
Age > 23 and Age <= 37
```

Impostazioni del nugget del modello Struttura ad albero delle decisioni/Insieme di regole

La scheda Impostazioni per una struttura da albero delle decisioni o un nugget del modello dell'insieme di regole consente di specificare le opzioni per le confidenze e la generazione del codice SQL durante il calcolo del punteggio del modello. Questa scheda è disponibile solo dopo che il nugget del modello è stato aggiunto a un flusso.

Calcola confidenze. Selezionare questa opzione per includere le confidenze nelle operazioni di calcolo del punteggio. Durante il calcolo del punteggio dei modelli nel database, l'esclusione delle confidenze consente di generare codice SQL più efficace. Per le strutture ad albero di regressione le confidenze non vengono assegnate.

Nota: se viene selezionata l'opzione **Crea un modello per insiemi di dati di grandi dimensioni** nella scheda Opzioni di creazione del pannello Metodo per i modelli CHAID, questa casella di controllo è disponibile solo nei nugget del modello per i target di categoria nominale o indicatore.

Calcola punteggi di propensione grezza. Per i modelli con obiettivi di tipo flag (che restituiscono un risultato sì o nessuna previsione), è possibile richiedere i punteggi di propensione grezza che indicano la verosimiglianza del risultato vero specificato per il campo obiettivo. Tali punteggi si aggiungono ai valori di previsione e confidenza che potrebbero venire generati durante il calcolo del punteggio.

Nota: se si seleziona l'opzione **Crea un modello per gli insiemi di dati di grandi dimensioni** nella scheda Opzioni di creazione del pannello Metodo per i modelli CHAID, questa casella di controllo è disponibile solo nei nugget del modello con target di categoria indicatore.

Calcola punteggi di propensione regolata. I punteggi di propensione grezza sono basati esclusivamente sui dati di addestramento e potrebbero essere eccessivamente ottimistici a causa della tendenza di molti modelli a sovradattare questi dati. Le propensioni regolate cercano di operare una compensazione esaminando le prestazioni del modello rispetto a una partizione di test o di convalida. Per utilizzare questa opzione è necessario che nel flusso sia definito un campo partizione e che nel nodo Modelli siano attivati i punteggi di propensione regolati, prima di generare il modello.

Nota: i punteggi di propensione regolata non sono disponibili per i modelli di struttura ad albero boosted e di insiemi di regole. Per ulteriori informazioni, consultare l'argomento “Modelli C5.0 boosted” a pagina 113.

ID regola. Per i modelli CHAID, QUEST e C&R Tree, questa opzione aggiunge un campo all'output del calcolo del punteggio che indica l'ID del nodo terminale al quale è assegnato ogni record.

Nota: quando è selezionata questa opzione, la generazione SQL non è disponibile.

Genera SQL per questo modello. Quando si utilizzano dati da un database, è possibile "restituire" codice SQL al database per l'esecuzione, migliorando in tal modo le performance di molte operazioni.

Selezionare una delle opzioni riportate di seguito per specificare il modo in cui viene eseguita la generazione del codice SQL.

- **Default: Calcola il punteggio mediante l'Adattatore per calcolo punteggio server (se installato), in caso contrario nel processo.** Se è stata eseguita la connessione ad un database con un adattatore per calcolo punteggio installato, genera il codice SQL utilizzando l'adattatore per calcolo punteggio, in caso contrario genera il codice SQL in modo nativo all'interno di SPSS Modeler.
- **Genera senza supporto per valori mancanti.** Selezionare questa opzione per attivare la generazione SQL senza l'overhead della gestione dei valori mancanti. Tale opzione imposta la previsione su null (`$null$`) quando viene trovato un valore mancante durante il calcolo del punteggio di un caso.

  Nota: questa opzione non è disponibile per i modelli CHAID. Per gli altri tipi di modelli, è disponibile solo per le strutture ad albero delle decisioni (non per gli insiemi di regole).
- **Genera con supporto per valori mancanti.** Per i modelli CHAID, QUEST e C&R Tree, è possibile abilitare la generazione del codice SQL con il supporto completo per valori mancanti. Ciò significa che viene generato SQL così i valori mancanti vengono gestiti come specificato nel modello. Ad esempio, C&R Trees utilizza regole surrogato e fallback del nodo figlio con più record.

  Nota: per i modelli C5.0, questa opzione è disponibile solo per gli insiemi di regole (non per le strutture ad albero delle decisioni).

Modelli C5.0 boosted

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

Quando si crea un modello C5.0 boosted (di un insieme di regole o di una struttura ad albero delle decisioni), in realtà viene creato un insieme di modelli correlati. Il browser delle regole del modello relativo a un modello C5.0 boosted mostra l'elenco di modelli al livello superiore della gerarchia, assieme alla precisione stimata di ogni modello e alla precisione complessiva dei modelli boosted dell'insieme. Per esaminare le regole o le suddivisioni relative a un determinato modello, selezionare tale modello ed espanderlo come se si trattasse di una regola o di un ramo in un modello singolo.

È inoltre possibile estrarre un determinato modello da un insieme di modelli boosted e creare un nuovo nugget del modello Insieme di regole contenente solo tale modello. Per creare un nuovo insieme di regole da un modello C5.0 boosted, selezionare l'insieme di regole o la struttura ad albero di interesse e scegliere tra **Struttura ad albero delle decisioni singola (palette MG)** oppure **Struttura ad albero delle decisioni singola (area)** dal menu Genera.

Generazione di grafici

I nodi Struttura ad albero forniscono una grande quantità di informazioni, che tuttavia non sempre sono in un formato facilmente accessibile per gli utenti di business. Per fornire dati che siano facilmente incorporabili nei report di business, nelle presentazioni, e così via, è possibile produrre dei grafici dei dati selezionati. Per esempio, dalla scheda Modello o Visualizzatore di un nugget del modello o dalla scheda Visualizzatore di una struttura ad albero interattiva, è possibile generare un grafico per una parte selezionata della struttura ad albero, creando quindi un grafico solo per i casi presenti nel nodo del ramo selezionato o della struttura ad albero selezionata.

Nota: è possibile generare un grafico da un nugget solo quando è collegato ad altri nodi in un flusso.

Generazione di un grafico

Come prima cosa, selezionare le informazioni che dovranno essere visualizzate nel grafico:

- Nella scheda Modello di un nugget, espandere l'elenco delle condizioni e delle regole nel riquadro di sinistra e selezionare l'elemento desiderato.
- Nella scheda Visualizzatore di un nugget, espandere l'elenco dei rami e selezionare il nodo desiderato.
- Nella scheda Visualizzatore di una struttura ad albero interattiva, espandere l'elenco dei rami e selezionare il nodo desiderato.

  Nota: non è possibile selezionare il primo nodo in alcuna delle schede Visualizzatore.

Il grafico viene creato allo stesso modo, a prescindere dal metodo di selezione dei dati da visualizzare:

1. Dal menu Genera, selezionare **Grafico (da selezione)** oppure, nella scheda Visualizzatore, fare clic sul pulsante **Grafico (da selezione)** nell'angolo inferiore sinistro. Viene visualizzata la scheda Di base del nodo Lavagna grafica.

   Nota: quando si visualizza la Lavagna grafica in questo modo, sono disponibili soltanto le schede Di base e Dettagliato.
2. Mediante le impostazioni della scheda Di base o Dettagliato, specificare i dettagli da visualizzare sul grafico.
3. Fare clic su OK per generare il grafico.

L'intestazione del grafico riporta i nodi o le regole scelte per l'inclusione.

Nugget del modello per boosting, bagging o insiemi di dati di grandi dimensioni

Se si seleziona **Ottimizza la precisione del modello (boosting)**, **Ottimizza la stabilità del modello (bagging)** o **Crea un modello per insiemi di dati di grandi dimensioni** come obiettivo principale nel nodo di modellazione, IBM SPSS Modeler crea un insieme di più modelli. Per ulteriori informazioni, consultare l'argomento "Modelli per insiemi" a pagina 45.

Il nugget del modello risultante conterrà le seguenti schede. La scheda Modello fornisce diverse visualizzazioni del modello.

Tabella 8. Schede disponibili nel nugget del modello

Scheda	Visualizza	Descrizione	Ulteriori informazioni
Modello	Riepilogo del modello	Visualizza un riepilogo della qualità e della diversità (eccetto per i modelli boosted e i target continui) dell'insieme, una misura di quanto le previsioni possono variare nei diversi modelli.	Per ulteriori informazioni, consultare l'argomento "Riepilogo del modello" a pagina 46.
	Importanza predittore	Visualizza un grafico che indica l'importanza relativa di ogni predittore (campo di input) nella stima del modello.	Per ulteriori informazioni, consultare l'argomento "Importanza predittore" a pagina 46.
	Frequenza dei predittori	Visualizza un grafico che indica l'importanza relativa con cui ogni predittore viene utilizzato nell'insieme di modelli.	Per ulteriori informazioni, consultare l'argomento "Frequenza dei predittori" a pagina 46.
	Precisione del modello di componente	Traccia un grafico della precisione predittiva di ciascun modello incluso nell'insieme.	
	Dettagli del modello di componente	Visualizza informazioni su ciascun modello incluso nell'insieme.	Per ulteriori informazioni, consultare l'argomento "Dettagli del modello di componenti" a pagina 47.
	Informazioni	Visualizza informazioni sui campi, le impostazioni di creazione e l'elaborazione della stima del modello.	Per ulteriori informazioni, consultare l'argomento "Scheda Riepilogo di un nugget del modello/Informazioni" a pagina 43.

Tabella 8. Schede disponibili nel nugget del modello (Continua)

Scheda	Visualizza	Descrizione	Ulteriori informazioni
Impostazioni		Consente di includere le confidenze nelle operazioni di calcolo del punteggio.	Per ulteriori informazioni, consultare l'argomento "Impostazioni del nugget del modello Struttura ad albero delle decisioni/Insieme di regole" a pagina 112.
Annotazione		Consente di aggiungere annotazioni descrittive, specificare un nome personalizzato, aggiungere un testo di suggerimento e inserire le parole chiave di ricerca per il modello.	

Nugget del modello dell'Insieme di regole

Un nugget del modello Insieme di regole rappresenta le regole per la previsione di un particolare campo di output rilevato dal nodo modellazione della regola di associazione (Apriori) o da uno dei nodi di creazione della struttura ad albero (C&R Tree, CHAID, QUEST o C5.0). Per le regole di associazione, l'insieme di regole deve essere generato da un nugget Regola grezza. Per le strutture ad albero, un insieme di regole può essere generato dal builder della struttura ad albero, da un nodo di creazione del modello C5.0 o da qualsiasi nugget del modello di struttura ad albero. A differenza dei nugget Regola grezza, è possibile posizionare i nugget Insieme di regole nei flussi per generare previsioni.

Quando viene eseguito un flusso che contiene un nugget Insieme di regole, al flusso vengono aggiunti due nuovi campi contenenti il valore previsto e la confidenza relativa a ogni record. Per il nome dei nuovi campi viene utilizzato il nome del modello con l'aggiunta di prefissi. Per le regole di associazione, i prefissi sono *$A-* per il campo della previsione e *$AC-* per il campo della confidenza. Per gli insiemi di regole C5.0, i prefissi sono *$C-* per il campo della previsione e *$CC-* per il campo della confidenza. Per gli insiemi di regole C&R Tree, i prefissi sono *$R-* per il campo della previsione e *$RC-* per il campo della confidenza. In un flusso composto da più nugget Insieme di regole in una serie che prevede gli stessi campi di output, i nomi dei nuovi campi includeranno numeri nel *prefisso* come criterio di differenziazione. Il primo nugget di modelli Insieme di regole di associazione nel flusso utilizzerà la denominazione standard, il secondo nodo aggiungerà i prefissi *$A1-* e *$AC1-*, il terzo i prefissi *$A2-* e *$AC2-* e così via.

Modalità di applicazione delle regole. Gli insiemi di regole generati da regole di associazione sono diversi da altri nugget del modello perché per ogni record è possibile generare più di una previsione e tali previsioni possono non essere tutte uguali. Sono disponibili due metodi per generare previsioni dagli insiemi di regole.

Nota: gli insiemi di regole generati da strutture ad albero delle decisioni restituiscono gli stessi risultati indipendentemente dal metodo utilizzato, poiché le regole derivate da una struttura ad albero delle decisioni si escludono a vicenda.

- **Confronto.** Questo metodo tenta di unire le previsioni di tutte le regole che si applicano al record. Per ogni record, vengono esaminate tutte le regole e ogni regola che si applica al record viene utilizzata per generare una previsione e una confidenza associata. Per ogni valore di output viene calcolata la somma delle cifre relative alla confidenza e come previsione finale viene scelto il valore con la somma di confidenza maggiore. La confidenza relativa alla previsione finale è la somma della confidenza relativa a tale valore diviso il numero di regole generate per tale record.
- **Primo risultato.** Questo metodo verifica le regole in ordine e la prima regola che si applica al record è quella utilizzata per generare la previsione.

È possibile controllare il metodo utilizzato nelle opzioni relative al flusso..

Generazione di nodi. Il menu Genera consente di creare nuovi nodi in base all'insieme di regole.

- **Nodo filtro.** Crea un nuovo nodo Filtro per filtrare i campi inutilizzati dalle regole nell'insieme di regole.
- **Nodo Seleziona.** Crea un nuovo nodo Selezione per selezionare i record a cui si applica la regola selezionata. Il nodo generato selezionerà i record in cui tutti gli antecedenti della regola sono veri. Questa opzione richiede la selezione di una regola.
- **Nodo Traccia regola.** Crea un nuovo Supernodo per il calcolo di un campo indicante la regola utilizzata per eseguire la previsione di ogni record. Quando viene valutato un insieme di regole utilizzando il metodo del primo risultato, verrà visualizzato un simbolo indicante la prima regola che verrà generata. Quando viene valutato l'insieme di regole utilizzando il metodo del confronto, verrà visualizzata una stringa più complessa che mostra l'input al meccanismo di confronto.
- **Struttura ad albero delle decisioni singola (area)/ Struttura ad albero delle decisioni singola (palette MG).** Crea un nuovo nugget Insieme di regole singolo derivato dalla regola correntemente selezionata. Disponibile solo nei modelli C5.0 **boosted**. Per ulteriori informazioni, consultare l'argomento "Modelli C5.0 boosted" a pagina 113.
- **Modello a palette.** Restituisce il modello nella palette dei modelli. Può rivelarsi utile in quelle situazioni in cui un collega abbia inviato un flusso che contiene il modello ma non il modello stesso.

Nota: le schede Impostazioni e Riepilogo del nugget Insieme di regole sono identiche a quelle utilizzate nei modelli di struttura ad albero delle decisioni.

Scheda Modello dell'Insieme di regole

La scheda Modello di un nugget Insieme di regole visualizza un elenco di regole che l'algoritmo estrae dai dati.

Le regole sono suddivise in base al valore consequent (categoria predittiva) e vengono presentate nel seguente formato:

```
if antecedent_1
and antecedent_2
...
and antecedent_n
then predicted value
```

dove `consequent` e i valori da `antecedent_1` ad `antecedent_n` sono tutte condizioni. La regola viene interpretata nel modo seguente: "nei record in cui i valori da `antecedent_1` a `antecedent_n` sono veri, è probabile che anche il valore `consequent` sia vero". Se si seleziona il pulsante **Mostra occorrenze/confidenza** nella barra degli strumenti, ogni regola visualizzerà anche informazioni relative al numero di record a cui si applica la regola, ovvero i record per cui sono veri gli antecedenti (**Istanze**) e la proporzione di tali record per cui è vera l'intera regola (**Confidenza**).

Si noti che la confidenza viene calcolata in modo leggermente diverso per gli insiemi di regole C5.0. C5.0 utilizza la formula seguente per il calcolo della confidenza di una regola:

```
(1 + number of records where rule is correct)
/
(2 + number of records for which the rule's antecedents are true)
```

Tale calcolo della stima di confidenza si adegua al processo di generalizzazione delle regole da una struttura ad albero delle decisioni (operazione eseguita da C5.0 quando viene creato un insieme di regole).

Importazione di progetti da AnswerTree 3.0

IBM SPSS Modeler è in grado di importare progetti salvati in AnswerTree 3.0 o 3.1 utilizzando la finestra di dialogo File > Apri standard, come riportato di seguito:

1. Dai menu di IBM SPSS Modeler scegliere:

   File > **Apri flusso**
2. Nell'elenco a discesa File di tipo selezionare **File di progetto AT (*.atp, *.ats)**.

   Ogni progetto importato viene convertito in un flusso di IBM SPSS Modeler contenente i nodi seguenti:

   - Un nodo origine che definisce la sorgente dati utilizzata, per esempio un file di dati IBM SPSS Statistics o una sorgente database.
   - Per ogni struttura ad albero nel progetto (ce ne possono essere varie), viene creato un nodo tipologia che definisce le proprietà per ogni campo (variabile), compreso tipo, ruolo (campo di input o predittore vs. campo di output o previsto), valori mancanti e altre opzioni.
   - Per ogni struttura ad albero nel progetto, vengono creati un nodo Partizione, che esegue la partizione dei dati per un campione di test o addestramento, e un nodo di creazione della struttura ad albero che definisce i parametri per la generazione della struttura ad albero (nodo C&R Tree, QUEST o CHAID).
3. Per visualizzare le strutture ad albero generate, eseguire il flusso.

Commenti

- Non è possibile esportare in AnswerTree le strutture ad albero delle decisioni generate in IBM SPSS Modeler. L'importazione può essere eseguita unicamente da AnswerTree a IBM SPSS Modeler.
- I profitti definiti in AnswerTree non vengono mantenuti quando il progetto viene importato in IBM SPSS Modeler.

Capitolo 7. Modelli di rete bayesiana

Nodo Rete bayesiana

Il nodo **Rete bayesiana** consente di generare un modello di probabilità combinando elementi osservati e registrati con conoscenze del mondo reale basate sul "buon senso" per stabilire la probabilità di occorrenze utilizzando attributi apparentemente non collegati fra loro. Il nodo si concentra sulle reti TAN (Tree Augmented Naïve Bayes) e Coperta di Markov principalmente utilizzate per la classificazione.

Le reti bayesiane vengono utilizzate per eseguire previsioni in situazioni molto diverse fra loro, per esempio:

- selezione di opportunità di prestito a basso rischio di insolvenza
- previsione del momento in cui determinate attrezzature necessiteranno di manutenzione, ricambi o sostituzione in base all'input di sensori e record esistenti
- risoluzione di problemi della clientela tramite strumenti di risoluzione dei problemi online
- diagnosi e risoluzione dei problemi di reti di telefonia mobile in tempo reale
- valutazione dei rischi e dei vantaggi potenziali di progetti di ricerca e sviluppo per orientare le risorse sulle opportunità migliori

Un rete bayesiana è un modello grafico che visualizza variabili (spesso definite **nodi**) in un dataset e le indipendenze probabilistiche o condizionali tra di esse. Le relazioni causali tra i nodi possono essere rappresentate da un rete bayesiana, tuttavia i collegamenti all'interno della rete (denominati anche **archi**) non rappresentano necessariamente una relazione causa-effetto diretta. Per esempio, è possibile utilizzare una rete bayesiana per calcolare la probabilità che un paziente soffra di una determinata malattia, data la presenza o l'assenza di certi sintomi e altri dati rilevanti, se le indipendenze probabilistiche tra i sintomi e la malattia visualizzate nel grafico si dimostrano vere. Le reti sono molto efficaci nei casi in cui mancano informazioni e forniscono la migliore previsione possibile utilizzando tutte le informazioni presenti.

Un esempio comune di rete bayesiana di base è quello creato da Lauritzen e Spiegelhalter nel 1988: spesso chiamato modello "Asia", è la versione semplificata di una rete utilizzabile per eseguire la diagnosi di nuovi pazienti, con la direzione dei collegamenti che corrisponde all'incirca alla causalità. Ogni nodo rappresenta un aspetto che potrebbe essere collegato alla condizione del paziente; per esempio "Smoking" indica che i pazienti sono fumatori dichiarati e "VisitAsia" mostra che si sono recati di recente in Asia. Le relazioni di probabilità sono mostrate dai collegamenti tra i nodi: per esempio, il fumo aumenta le probabilità che il paziente soffra di bronchite e cancro al polmone, mentre l'età sembra collegata solo alla possibilità di sviluppare il cancro al polmone. Allo stesso modo, le anomalie di una radiografia ai polmoni possono essere causate da tubercolosi o cancro ai polmoni, mentre le probabilità che un paziente soffra di dispnea sono maggiori se il paziente soffre anche di bronchite o cancro al polmone.

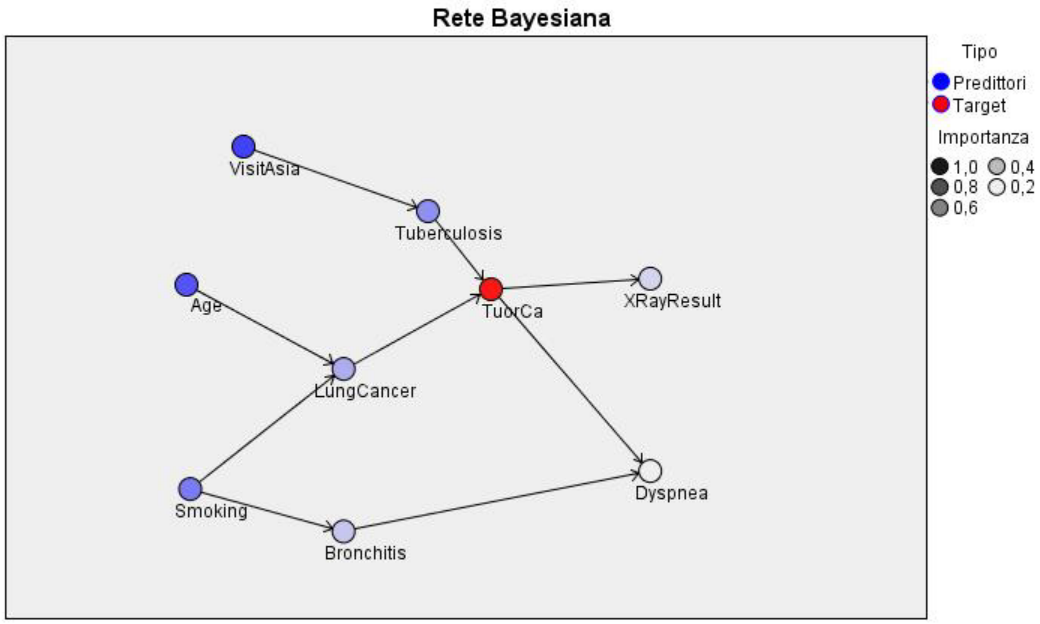


Figura 29. Esempio di rete Asia di Lauritzen e Spiegelhalter

È possibile decidere di ricorrere a una rete bayesiana per svariate ragioni:

- facilita la comprensione delle relazioni casuali consentendo di capire un'area problematica e di prevedere le conseguenze di un eventuale intervento
- la rete costituisce un approccio efficiente per evitare il sovradattamento di dati
- è facile osservare una chiara visualizzazione delle relazioni coinvolte

Requisiti. I campi obiettivo devono essere categoriali e possono avere un livello di misurazione *Nominale*, *Ordinale* o *Flag*. Gli input possono essere campi di qualsiasi tipo. I campi di input continui (intervallo numerico) vengono automaticamente raccolti; tuttavia, se la distribuzione è asimmetrica, è possibile ottenere risultati migliori eseguendo manualmente la discretizzazione anteponendo un nodo Discretizza al nodo Rete bayesiana. Per esempio, utilizzare la discretizzazione ottimale dove il **campo supervisore** corrisponde al campo **Obiettivo** del nodo Rete bayesiana.

Esempio. L'analista di una banca desidera poter prevedere quali clienti o potenziali clienti non restituiranno probabilmente i prestiti ricevuti. È possibile utilizzare un modello di rete bayesiana per identificare le caratteristiche dei clienti con maggiori probabilità di essere insolventi e creare diversi tipi di modelli differenti per stabilire il migliore per la previsione di potenziali clienti insolventi.

Esempio. Un operatore telefonico vuole ridurre il numero di clienti che abbandonano l'azienda (detto "tasso di abbandono") e aggiornare il modello mensilmente utilizzando i dati di ogni mese precedente. È possibile utilizzare un modello di rete bayesiana per identificare le caratteristiche dei clienti con maggior probabilità di abbandonare e continuare ad addestrare il modello ogni mese con nuovi dati.

Opzioni Modello di un nodo Rete bayesiana

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.

Utilizza dati partizionati. Se è definito un campo partizione, questa opzione garantisce che per la creazione del modello vengano utilizzati dati solo dalla partizione di addestramento.

Crea un modello per ciascuna suddivisione. Crea un modello separato per ogni valore possibile dei campi di input che sono stati specificati come campi di suddivisione. Per ulteriori informazioni, consultare l'argomento "Creazione di modelli di suddivisione" a pagina 28.

Partizione. Questo campo consente di specificare un campo utilizzato per partizionare i dati in campioni separati per le 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 un'indicazione valida del modo in cui il modello potrà essere esteso a insiemi di dati di dimensioni maggiori, simili ai dati correnti. Se sono stati definiti più campi partizione utilizzando nodi Tipo o Partizione, 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. Si noti inoltre che per applicare la partizione selezionata nell'analisi, è necessario che il partizionamento sia attivato anche nella scheda Opzioni modello relativa al nodo. La disattivazione di questa opzione consente di disattivare il partizionamento senza dover modificare le impostazioni dei campi.

Suddivisioni. Per i modelli di suddivisione, selezionare il campo o i campi di suddivisione. Questa operazione è simile all'impostazione del ruolo di un campo su *Suddivisione* in un nodo tipologia. È possibile designare come campi di suddivisione solo i campi con livello di misurazione **Flag**, **Nominale**, **Ordinale** o **Continuo**. I campi selezionati come campi di suddivisione non possono essere utilizzati come campi obiettivo, di input, di partizione, di frequenza o peso. Per ulteriori informazioni, consultare l'argomento "Creazione di modelli di suddivisione" a pagina 28.

Addestramento continuo modello esistente. Se si seleziona questa opzione, i risultati riportati nella scheda Modello del nugget del modello vengono rigenerati e aggiornati ogni volta che si esegue il modello. Per esempio, si eseguirà questa operazione quando sono state aggiunte sorgenti dati nuove o aggiornate a un modello esistente.

Nota: questa opzione consente solo di aggiornare la rete esistente; non è in grado di aggiungere o rimuovere nodi o connessioni. Ogni volta che si riaddestra il modello, la rete manterrà la stessa forma, mentre verranno modificate solo le probabilità condizionali e l'importanza dei predittori. Se i nuovi dati sono complessivamente simili a quelli vecchi, questo non pone problemi poiché si prevede che siano significativi gli stessi elementi. Tuttavia se si desidera verificare o aggiornare *che cosa* è significativo (anziché quanto importante esso sia), sarà necessario costruire un nuovo modello, ovvero una nuova rete.

Tipo di struttura. Selezionare la struttura da utilizzare quando si costruisce la rete bayesiana:

- **TAN.** Il modello TAN (Tree Augmented Naïve Bayes) crea un modello di rete bayesiana semplice che costituisce un miglioramento del modello Naïve Bayes standard. Questo perché consente a ogni predittore di dipendere da un altro predittore in aggiunta alla variabile obiettivo, aumentando pertanto la precisione della classificazione.
- **Coperta di Markov.** Questa opzione seleziona l'insieme dei nodi nel dataset che contengono gli elementi della variabile obiettivo, i relativi elementi figlio e gli elementi padre dei relativi elementi figlio. In pratica, una rete coperta di Markov identifica tutte le variabili della rete necessarie per prevedere la variabile obiettivo. Questo modo di generare una rete è considerato più preciso; tuttavia, con insiemi di dati di grandi dimensioni potrebbe richiedere un tempo di elaborazione superiore considerato il numero elevato di variabili coinvolte. Per ridurre i tempi di elaborazione, è possibile utilizzare le opzioni **Selezione funzioni** della scheda Livello avanzato per selezionare le variabili che sono significativamente correlate alla variabile obiettivo.

Includi procedura preliminare all'elaborazione di selezione funzioni. Se si seleziona questa casella, è possibile utilizzare le opzioni **Selezione funzioni** della scheda Livello avanzato.

Metodo di apprendimento parametro. I parametri delle reti bayesiane si riferiscono alle probabilità condizionali di ogni nodo, dati i valori dei suoi elementi padre. Sono disponibili due possibili opzioni per controllare l'attività di stima delle tabelle di probabilità condizionale tra nodi laddove i valori degli elementi padre sono noti:

- **Massima verosimiglianza.** Selezionare questa casella quando si usa un dataset di grandi dimensioni. Questa è l'opzione di default.
- **Correzione di Bayes per conteggi di celle piccoli.** Per insiemi di dati di dimensioni più piccole sussiste il pericolo di un sovradattamento del modello, nonché la possibilità di un numero elevato di conteggi zero. Selezionare questa opzione per limitare questi problemi applicando il livellamento per ridurre l'effetto di eventuali conteggi zero e gli effetti di stime non affidabili.

Opzioni avanzate del nodo Rete bayesiana

Le opzioni avanzate del nodo consentono di definire con precisione il processo di creazione dei modelli. Per accedere alle opzioni avanzate, impostare Moda su **Livello avanzato** nella scheda Livello avanzato.

Valori mancanti. Per default, IBM SPSS Modeler utilizza solo i record con valori validi per tutti i campi utilizzati nel modello (questo approccio viene a volte chiamato **eliminazione listwise** dei valori mancanti). Tuttavia, se la quantità di valori mancanti è notevole, è possibile che in questo modo venga eliminato un numero eccessivo di record e che pertanto i dati disponibili non siano sufficienti per generare un modello efficace. In tali casi, è possibile deselezionare l'opzione **Utilizza solo record completi**. IBM SPSS Modeler quindi prova ad utilizzare la maggior quantità di informazioni possibile per eseguire la stima del modello, inclusi i record in cui alcuni dei campi hanno valori mancanti. (questo approccio viene a volte chiamato **eliminazione pairwise** dei valori mancanti). Tuttavia, in alcune situazioni, l'utilizzo di record incompleti può portare a problemi di calcolo nella stima del modello.

Accoda tutte le probabilità. Specifica se le probabilità di ogni categoria del campo di output vengono aggiunte a ogni record elaborato dal nodo. Se questa opzione non è selezionata, verrà aggiunta solo la probabilità della categoria prevista.

Test di indipendenza. Un test di indipendenza valuta se le osservazioni accoppiate su due variabili sono indipendenti tra loro. Selezionare il tipo di test da utilizzare tra le seguenti opzioni:

- **Rapporto di verosimiglianza.** Verifica l'indipendenza obiettivo-predittore calcolando un rapporto tra la massima probabilità di un risultato secondo due diverse ipotesi.
- **Chi-quadrato di Pearson.** Verifica l'indipendenza obiettivo-predittore utilizzando l'ipotesi nulla secondo cui le frequenze relative di occorrenza di eventi osservati seguono una specifica distribuzione della frequenza.

I modelli di rete bayesiana eseguono test condizionali di indipendenza dove, oltre alle coppie testate, vengono utilizzate ulteriori variabili. Inoltre, i modelli non si limitano a esplorare le relazioni tra obiettivo e predittori, ma anche le relazioni tra i predittori stessi.

Nota: le opzioni Test di indipendenza sono disponibili solo se si seleziona **Includi procedura preliminare all'elaborazione di selezione funzioni** oppure un **Tipo di struttura** coperta di Markov nella scheda Modello.

Livello di significatività. Opzione utilizzata in combinazione con le impostazioni Test di indipendenza, che consente di impostare un valore di interruzione da utilizzare quando si effettuano i test. Minore è il valore, minore è il numero di collegamenti che rimangono nella rete. Il livello di default è 0.01.

Nota: questa opzione è disponibile solo se si seleziona **Includi procedura preliminare all'elaborazione di selezione funzioni** oppure un **Tipo di struttura** coperta di Markov nella scheda Modello.

Dimensione insieme di condizionamento massima. L'algoritmo per la creazione di una struttura coperta di Markov utilizza insiemi di condizionamento di dimensioni crescenti per eseguire test di indipendenza

e rimuovere collegamenti non necessari dalla rete. Poiché i test che coinvolgono un numero elevato di variabili di condizionamento richiedono più tempo e memoria per l'elaborazione è possibile limitare il numero di variabili da includere. Questo può essere particolarmente utile quando si elaborano dati con forti dipendenze tra molte variabili. Si noti tuttavia che la rete risultante potrebbe contenere alcuni collegamenti superflui.

Specificare il numero massimo di variabili di condizionamento da utilizzare per il test di indipendenza. L'impostazione predefinita è 5.

Nota: questa opzione è disponibile solo se si seleziona **Includi procedura preliminare all'elaborazione di selezione funzioni** oppure un **Tipo di struttura** coperta di Markov nella scheda Modello.

Selezione funzioni. Queste opzioni consentono di limitare il numero di input utilizzati durante l'elaborazione del modello per accelerarne il processo di creazione. Questa opzione è particolarmente utile quando si crea una struttura coperta di Markov, a causa del possibile numero elevato di input potenziali. Consente infatti di selezionare gli input che sono significativamente correlati alla variabile obiettivo.

Nota: le opzioni di selezione delle funzioni sono disponibili solo se si seleziona **Includi procedura preliminare all'elaborazione di selezione funzioni** nella scheda Modello.

- **Input sempre selezionati** Con il selettore di campo (pulsante situato a destra del campo testo), selezionare i campi del dataset che devono essere sempre utilizzati quando si genera il modello di rete bayesiana. Si noti che il campo obiettivo è sempre selezionato.
- **Numero max di input.** Specificare il numero totale di input del dataset da utilizzare quando si genera il modello di rete bayesiana. Il numero massimo che è possibile immettere è il numero totale di input presente nel dataset.

  Nota: se il numero di campi selezionati in **Input sempre selezionati** supera il valore di **Numero max di input**, viene visualizzato un messaggio di errore.

Nugget del modello di rete bayesiana

Nota: se è stata selezionata l'opzione **Addestramento continuo parametri esistenti** nella scheda Modello del nodo modellazione, le informazioni visualizzate nella scheda Modello del nugget del modello vengono aggiornate ogni volta che il modello viene rigenerato.

La scheda Modello del nugget del modello è suddivisa in due riquadri:

Riquadro sinistro

Di base. Questa visualizzazione contiene una rete di nodi che mostra la relazione tra l'obiettivo e i suoi predittori più importanti, nonché la relazione tra i predittori. L'importanza di ogni predittore è mostrata dalla densità del suo colore, con un colore intenso che indica un predittore importante e viceversa.

I valori bin dei nodi che rappresentano un intervallo vengono visualizzati in una finestra descrittiva a comparsa quando si passa il puntatore del mouse sul nodo.

Per interagire con il grafico, modificarlo e salvarlo è possibile utilizzare gli strumenti per i grafici di IBM SPSS Modeler, per esempio per l'utilizzo in altre applicazioni come MS Word.

Suggerimento: se la rete contiene un numero elevato di nodi, è possibile fare clic su un nodo e trascinarlo in modo da rendere il grafico più leggibile.

Distribuzione. Questa visualizzazione mostra le probabilità condizionali per ogni nodo della rete sotto forma di mini grafico. Passare il puntatore del mouse su un grafico per visualizzarne i valori in una finestra descrittiva a comparsa.

Riquadro destro

Importanza predittore. Visualizza un grafico che indica l'importanza relativa di ogni predittore nella stima del modello. Per ulteriori informazioni, consultare l'argomento “Importanza predittore” a pagina 43.

Probabilità condizionali. Quando si seleziona un nodo o un mini grafico di distribuzione nel riquadro sinistro, la tabella delle probabilità condizionali associate viene visualizzata nel riquadro destro. Questa tabella contiene il valore di probabilità condizionale per ogni valore del nodo e ogni combinazione di valori nei corrispondenti nodi padre. Include inoltre il numero di record osservati per ogni valore di record e ogni combinazione di valori nei nodi padre.

Impostazioni dei modelli di rete bayesiana

La scheda Impostazioni di un nugget del modello Rete bayesiana contiene le opzioni per la modifica del modello creato. Per esempio, è possibile utilizzare il nodo Rete bayesiana per creare vari modelli diversi a partire dagli stessi dati e dalle stesse impostazioni, quindi utilizzare questa scheda in ogni modello per modificare leggermente le impostazioni al fine di verificare in che modo questo influisca sui risultati.

Nota: questa scheda è disponibile solo dopo che il nugget del modello è stato aggiunto a un flusso.

Calcola punteggi di propensione grezza. Per i modelli con obiettivi di tipo flag (che restituiscono un risultato sì o nessuna previsione), è possibile richiedere i punteggi di propensione grezza che indicano la verosimiglianza del risultato vero specificato per il campo obiettivo. Tali punteggi si aggiungono ai valori di previsione e confidenza che potrebbero venire generati durante il calcolo del punteggio.

Calcola punteggi di propensione regolata. I punteggi di propensione grezza sono basati esclusivamente sui dati di addestramento e potrebbero essere eccessivamente ottimistici a causa della tendenza di molti modelli a sovradattare questi dati. Le propensioni regolate cercano di operare una compensazione esaminando le prestazioni del modello rispetto a una partizione di test o di convalida. Per utilizzare questa opzione è necessario che nel flusso sia definito un campo partizione e che nel nodo Modelli siano attivati i punteggi di propensione regolati, prima di generare il modello.

Accoda tutte le probabilità. Specifica se le probabilità di ogni categoria del campo di output vengono aggiunte a ogni record elaborato dal nodo. Se questa opzione non è selezionata, verrà aggiunta solo la probabilità della categoria prevista.

L'impostazione di default di questa casella di controllo è determinata dalla corrispondente casella di controllo nella scheda Livello avanzato del nodo Modelli. Per ulteriori informazioni, consultare l'argomento “Opzioni avanzate del nodo Rete bayesiana” a pagina 122.

Riepilogo di un modello Rete bayesiana

La scheda Riepilogo di un nugget del modello visualizza le informazioni sul modello stesso (*Analisi*), i campi utilizzati nel modello (*Campi*), le impostazioni utilizzate durante la creazione del modello (*Impostazioni di creazione*) e l'addestramento del modello (*Riepilogo addestramento*).

Quando si sfoglia per la prima volta il nodo, i risultati della scheda Riepilogo sono compressi. Per visualizzare i risultati, utilizzare il controllo dell'espansore a sinistra di un elemento se si desidera visualizzare solo i risultati di tale elemento oppure fare clic sul pulsante **Espandi tutto** se si desidera visualizzare tutti i risultati. Per nascondere i risultati una volta terminata la visualizzazione, utilizzare il controllo dell'espansore per comprimere i risultati specifici che si desidera nascondere oppure fare clic sul pulsante **Comprimi tutto** per comprimere tutti i risultati.

Analisi. Visualizza informazioni sul modello specifico.

Campi. Elenca i campi utilizzati come obiettivi e come input nella creazione del modello.

Impostazioni di creazione. Contiene informazioni sulle impostazioni utilizzate nella creazione del modello.

Riepilogo addestramento. Mostra il tipo di modello, il flusso utilizzato per creare tale modello, l'utente che lo ha creato, la data di creazione e il tempo trascorso per la creazione del modello.

Capitolo 8. Reti neurali

Una **rete neurale** è in grado di approssimare un'ampia gamma di modelli predittivi con un impegno minimo del modello in termini di struttura e presupposti. La forma delle relazioni viene determinata durante il processo di apprendimento. Se una relazione lineare tra l'obiettivo e i predittori è appropriata, i risultati della rete neurale devono avvicinarsi molto a quelli di un modello lineare tradizionale. Se una relazione non lineare risulta più appropriata, la rete neurale approssima automaticamente la struttura "corretta" del modello.

Il compromesso per tale flessibilità è il fatto che la rete neurale non è facilmente interpretabile. Se è necessario spiegare qual è il processo sottostante che genera la relazione tra l'obiettivo e i predittori, è preferibile utilizzare un modello statistico più tradizionale. Tuttavia, se l'interpretazione del modello non è importante, è possibile ottenere previsioni apprezzabili con una rete neurale.

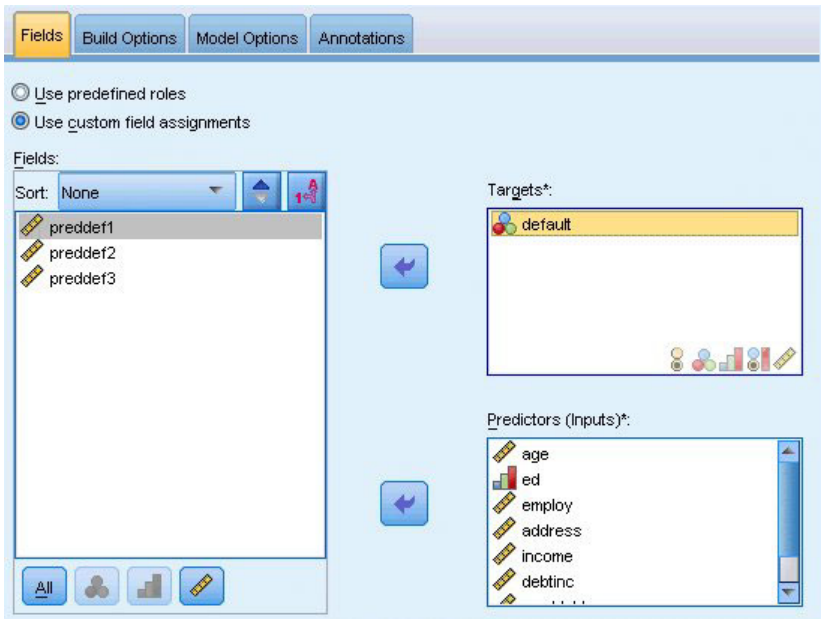


Figura 30. Scheda Campi

Requisiti dei campi. Devono essere presenti almeno un obiettivo ed un input. I campi impostati su Entrambi o Nessuno verranno ignorati. Non esistono restrizioni per il livello di misurazione di obiettivi o predittori (input). Per ulteriori informazioni, consultare l'argomento “Opzioni dei campi dei nodi Modelli” a pagina 31.

Modello di reti neurali

Le reti neurali sono semplici modelli del modo in cui opera il sistema nervoso. Le unità di base sono costituite dai **neuroni**, in genere organizzati in **strati**, come illustrato nella figura riportata di seguito.

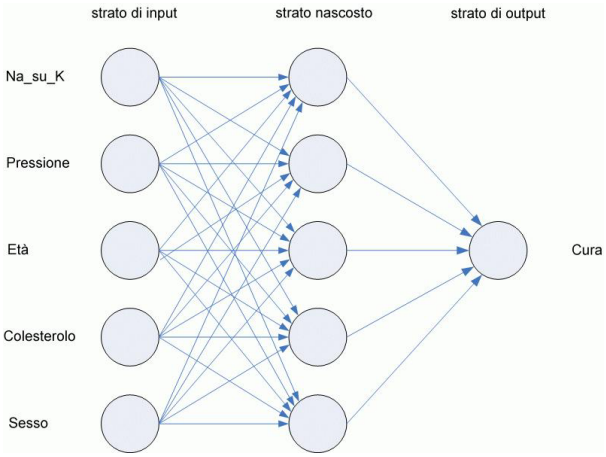


Figura 31. Struttura di una rete neurale

Una **rete neurale** è un modello semplificato del modo in cui il cervello umano elabora le informazioni. Funziona simulando un elevato numero di unità di elaborazione interconnesse che assomigliano a versioni astratte di neuroni.

Le unità di elaborazione sono disposte in strati. In una rete neurale sono generalmente presenti tre parti: uno **strato di input**, con unità che rappresentano i campi di input, uno o più **strati nascosti** ed uno **strato di output**, con una o più unità che rappresentano i campi obiettivo. Le unità sono connesse con varie intensità di connessione (o **pesi**). I dati di input vengono presentati nel primo strato e i valori vengono propagati da ciascun neurone a ogni neurone nello strato successivo. Alla fine viene fornito un risultato nello strato di output.

La rete apprende esaminando i singoli record, generando una previsione per ciascuno di essi e correggendo i pesi ogni volta che sbaglia previsione. Questo processo viene ripetuto molte volte e la rete continua a migliorare le sue previsioni finché uno o più criteri di interruzione non vengono soddisfatti.

Tutti i pesi inizialmente sono casuali e le risposte ottenute dalla rete risulteranno probabilmente prive di senso. L'impostazione della rete viene completata mediante l'**apprendimento**. Alla rete vengono presentati ripetutamente esempi con output conosciuti e le risposte fornite vengono confrontate con i risultati noti. Tramite la rete vengono quindi restituite le informazioni ottenute da questo confronto, modificando gradualmente i pesi. Nel corso del processo di apprendimento, la replica dei risultati noti da parte della rete risulta sempre più precisa. Al termine dell'apprendimento, la rete può essere applicata a casi futuri per cui non sono disponibili risultati conosciuti.

Utilizzo delle reti neurali con i flussi di versioni precedenti

Nella versione 14 di IBM SPSS Modeler è disponibile un nuovo nodo Rete neurale con il supporto delle tecniche di boosting e bagging e l'ottimizzazione per gli insiemi di dati di grandi dimensioni. I flussi già esistenti che contengono il tipo di nodo precedente vengono mantenuti in questa versione per la creazione e il calcolo del punteggio dei modelli. Tuttavia, tale supporto verrà rimosso nelle versioni future, quindi si consiglia di utilizzare la nuova versione a partire da questo momento.

A partire dalla versione 13, i campi con valori sconosciuti (valori non presenti nei dati di addestramento) non vengono più considerati automaticamente valori mancanti e il relativo punteggio viene calcolato con il valore `$null$`. Pertanto, se si desidera calcolare il punteggio di campi con valori sconosciuti come non null utilizzando un modello di Rete neurale precedente (pre-13) nella versione 13 o successiva, è necessario contrassegnare i valori sconosciuti come valori mancanti (per esempio, tramite il nodo Tipo).

Notare che, per motivi di compatibilità, qualsiasi flusso legacy che ancora contiene il nodo precedente può ancora utilizzare l'opzione *Limita dimensione insieme* da **Strumenti > Proprietà flusso > Opzione**; questa opzione è valida solo per le reti Kohonen ed i nodi *K*-medie a partire dalla versione 14.

Obiettivi

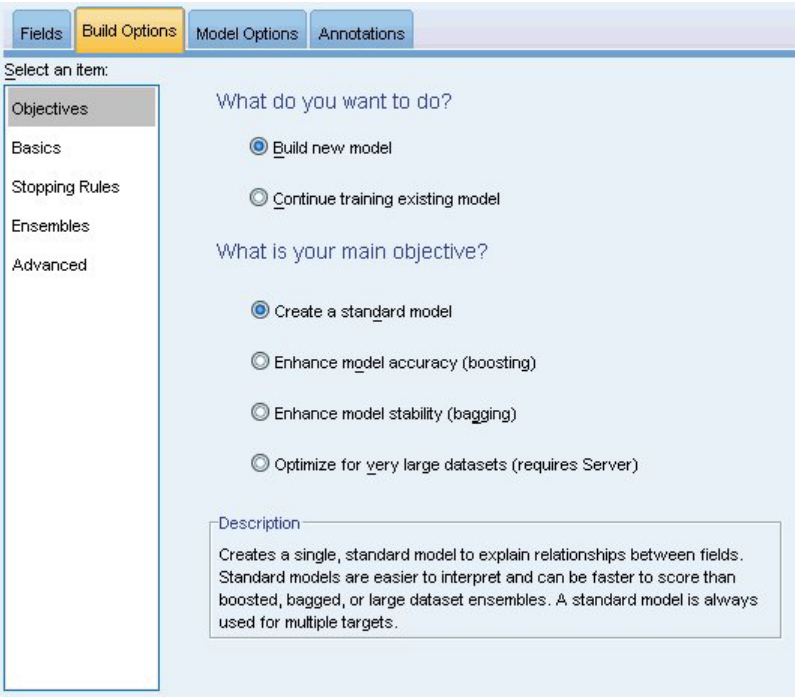


Figura 32. Impostazioni degli obiettivi

Come si desidera procedere?

- **Creare un nuovo modello.** Creare un modello completamente nuovo. Questa è la procedura normale per il nodo.
- **Continuare l'addestramento di un modello già esistente.** L'addestramento continua con l'ultimo modello generato correttamente dal nodo. In questo modo sarà possibile aggiornare un modello esistente senza dover accedere ai dati originali, aumentando così significativamente le prestazioni poiché nel flusso verranno utilizzati solo i record nuovi o aggiornati. I dettagli del modello precedente vengono archiviati con il nodo Modelli, consentendo di utilizzare questa opzione anche se il nugget del modello precedente non è più disponibile nel flusso o nella palette Modelli.

  Nota: quando questa opzione è abilitata, tutti gli altri controlli nelle schede Opzioni di creazione e dei campi sono disabilitati.

Qual è l'obiettivo principale? Selezionare l'obiettivo appropriato.

- **Creare un modello standard.** Il metodo crea un unico modello per la previsione dell'obiettivo utilizzando i predittori. In generale, i modelli standard sono più semplici da interpretare e consentono un più rapido calcolo del punteggio rispetto agli insiemi utilizzati per boosting, bagging o insiemi di dati di grandi dimensioni.
- **Migliorare l'accuratezza del modello (boosting).** Il metodo crea un modello di classificazione binario mediante la tecnica di boosting, che genera una sequenza di modelli per ottenere previsioni più precise. La creazione e il calcolo del punteggio degli insiemi può richiedere più tempo rispetto al modello standard.

  Il boosting genera una successione di "modelli di componenti", ciascuno dei quali viene creato a partire dall'intero insieme di dati. Prima di creare ogni successivo modello di componente, i record vengono

ponderati in base ai residui del modello di componente precedente. Ai casi con molti residui viene attribuito un peso di analisi relativamente superiore per far sì che il modello di componente successivo si concentri sulla corretta previsione di quei record. Insieme, questi modelli di componenti formano un modello di classificazione binario. Tale modello calcola il punteggio dei nuovi record utilizzando una regola di combinazione; le regole disponibili dipendono dal livello di misurazione dell'obiettivo.

- **Migliorare la stabilità del modello (bagging).** Il metodo crea un modello di classificazione binario mediante la tecnica di bagging (bootstrap aggregating), che genera più modelli per ottenere previsioni più affidabili. La creazione e il calcolo del punteggio degli insiemi può richiedere più tempo rispetto al modello standard.

  L'aggregazione bootstrap (bagging) genera repliche dell'insieme di dati di addestramento mediante il campionamento con sostituzione dall'insieme di dati originale. Questa operazione crea campioni di bootstrap di dimensioni uguali a quelle dell'insieme di dati originale. Su ogni replica viene quindi creato un "modello di componente". Insieme, questi modelli di componenti formano un modello di classificazione binario. Tale modello calcola il punteggio dei nuovi record utilizzando una regola di combinazione; le regole disponibili dipendono dal livello di misurazione dell'obiettivo.
- **Creare un modello per dataset di grandi dimensioni (richiede IBM SPSS Modeler Server).** Il metodo crea un modello di classificazione binario suddividendo l'insieme di dati in blocchi di dati separati. Scegliere questa opzione se l'insieme di dati è troppo grande per creare uno dei modelli descritti sopra o per la creazione incrementale del modello. Questa opzione può richiedere meno tempo per la creazione, ma più tempo per il calcolo del punteggio rispetto al modello standard. Questa opzione richiede la connettività IBM SPSS Modeler Server .

Se vi sono più obiettivi, questo metodo crea solo un modello standard, indipendentemente dall'obiettivo selezionato.

Impostazioni di base

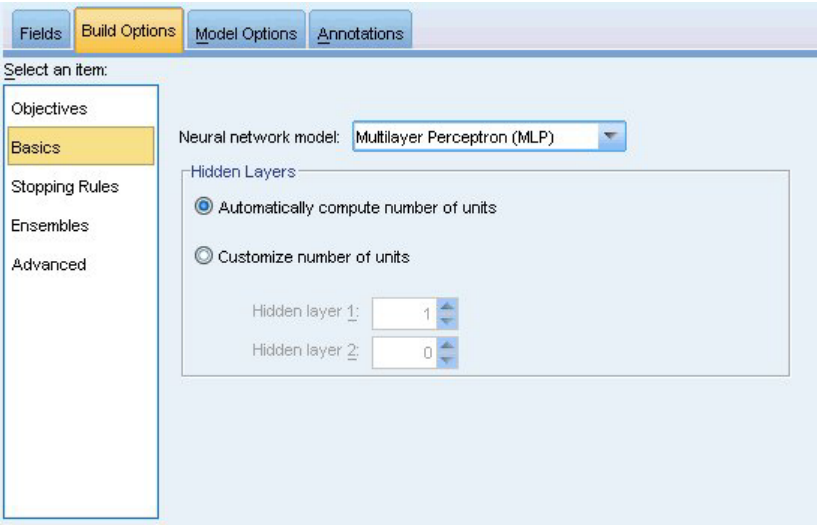


Figura 33. Impostazioni di base

Modello rete neurale. Il tipo di modello determina il modo in cui la rete collega i predittori agli obiettivi attraverso gli strati nascosti. L'algoritmo **MLP (percettore multistrato)** consente la creazione di relazioni più complesse che, però, implicano un aumento dei tempi di addestramento e di calcolo del punteggio. L'algoritmo **Funzione a base radiale (RBF)** può richiedere tempi più brevi per l'addestramento e il calcolo del punteggio, ma offre un potere predittivo ridotto rispetto a MLP.

Strati nascosti. Gli strati nascosti di una rete neurale sono costituiti da unità non osservabili. Il valore di ogni unità nascosta è una funzione dei predittori, la cui forma esatta dipende in parte dal tipo di rete. Un percettore multistrato può avere uno o due strati nascosti, una rete RBF può avere un solo strato nascosto.

- **Calcola automaticamente il numero di unità.** Questa opzione consente di creare una rete con uno strato nascosto e calcola il numero "migliore" di unità presenti nello strato nascosto.
- **Personalizza il numero di unità.** Questa opzione consente di specificare il numero di unità in ogni strato nascosto. Il primo strato nascosto deve contenere almeno un'unità. Se vengono specificate 0 unità per il secondo strato nascosto, viene creato un percettore multistrato con un solo strato nascosto.

  Nota: è necessario scegliere i valori in modo che il numero di nodi non sia maggiore della somma del numero di predittori continui e del numero totale di categorie tra tutti i predittori relativi alla categoria (flag, nominali ed ordinali).

Regole di arresto

Figura 34. Impostazioni delle regole di arresto

Le regole di arresto determinano quando interrompere l'addestramento delle reti MLP e vengono ignorate se si utilizza l'algoritmo RBF. L'addestramento procede per almeno un ciclo (passaggio di dati) e successivamente può essere arrestato in base ai seguenti criteri.

Utilizza tempo di addestramento massimo (per modello di componente). Scegliere se specificare un numero massimo di minuti per l'algoritmo da eseguire. Specificare un numero maggiore di 0. Se si crea un modello di classificazione binario, questo è il tempo di addestramento consentito per ciascun modello componente dell'insieme. Si noti che, per completare il ciclo corrente, l'addestramento potrebbe richiedere un tempo leggermente superiore al limite specificato.

Personalizza il numero massimo di cicli di addestramento. Numero massimo di cicli di addestramento consentiti. Se viene superato il numero massimo di cicli, l'addestramento si interrompe. Specificare un intero maggiore di 0.

Utilizza precisione minima. Con questa opzione, l'addestramento continuerà finché non viene raggiunta la precisione specificata. Questa condizione potrebbe non verificarsi mai, ma è possibile interrompere l'addestramento in qualsiasi punto e salvare la rete con la migliore precisione raggiunta fino a quel momento.

L'algoritmo di addestramento si interrompe anche se l'errore dell'insieme di prevenzione del sovradattamento non diminuisce dopo ogni ciclo, se la variazione relativa nell'errore di addestramento è minima o se il rapporto dell'errore di addestramento corrente è ridotto rispetto all'errore iniziale.

Insiemi

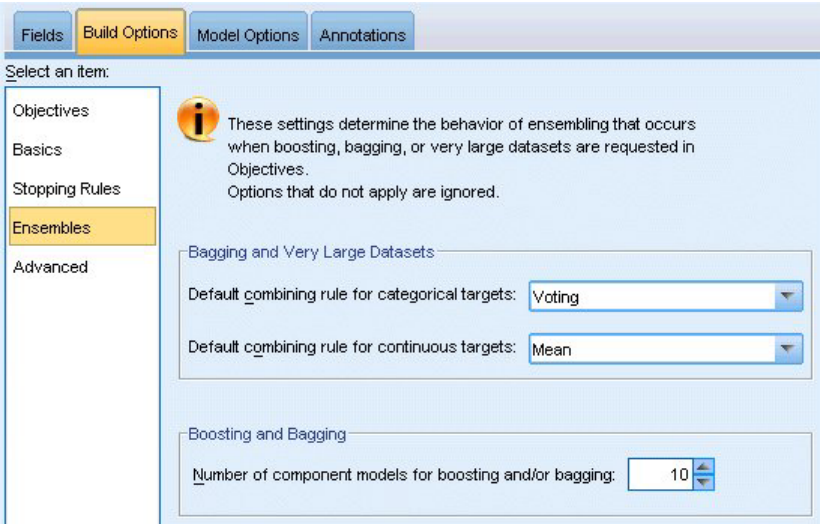


Figura 35. Impostazioni degli insiemi

Queste impostazioni determinano il comportamento dei classificatori binari che si verificano quando negli obiettivi sono richiesti boosting, bagging o insiemi di dati di grandi dimensioni. Le opzioni non applicate all'obiettivo selezionato vengono ignorate.

Bagging e insiemi di dati di grandi dimensioni. Quando si calcola il punteggio di un insieme, questo tipo di regola consente di combinare i valori previsti provenienti dai modelli di base per calcolare il valore del punteggio dell'insieme.

- **Regola di combinazione di default per gli obiettivi categoriali.** I valori previsti dell'insieme per gli obiettivi categoriali possono essere combinati utilizzando il confronto, la probabilità massima o la probabilità media più elevata. La **votazione** consente di selezionare la categoria che presenta più spesso la probabilità più elevata nei modelli di base. La **probabilità più elevata** consente di selezionare la categoria che raggiunge la singola probabilità più elevata in tutti i modelli di base. La **probabilità media più elevata** consente di selezionare la categoria con il massimo valore quando viene calcolata la media delle probabilità della categoria in tutti i modelli di base.
- **Regola di combinazione di default per target continui.** I valori previsti degli insiemi per i target continui possono essere combinati utilizzando la media o la mediana dei valori previsti ricavati dai modelli di base.

Si noti che quando l'obiettivo è di ottimizzare la precisione del modello, le selezioni delle regole di combinazione vengono ignorate. Nel boosting viene sempre utilizzato un voto di maggiore ponderazione per il calcolo del punteggio degli obiettivi categoriali e una mediana pesata per il calcolo del punteggio dei target continui.

Boosting e bagging. Specificare il numero dei modelli di base da creare quando l'obiettivo è di ottimizzare la precisione o la stabilità del modello. Per il bagging, si tratta del numero di campioni di bootstrap. Questo valore deve essere un numero intero positivo.

Impostazioni avanzate

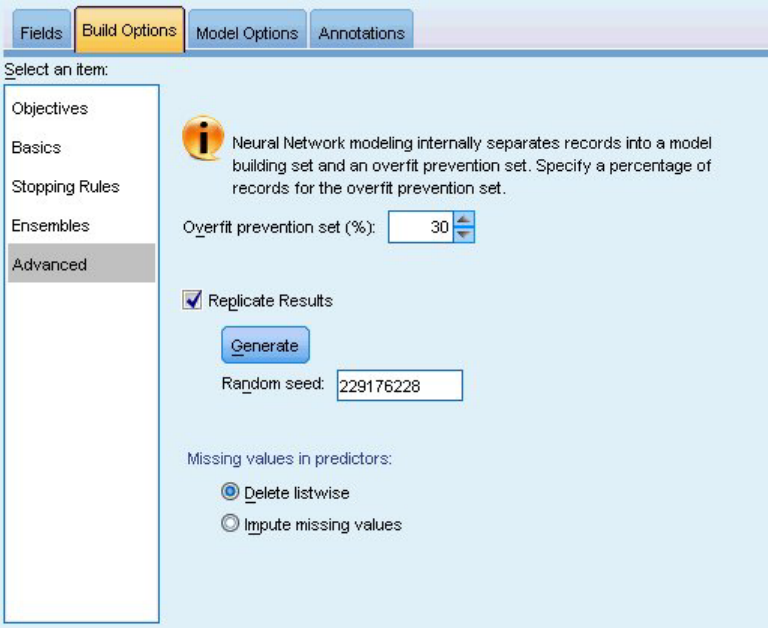


Figura 36. Impostazioni avanzate

Le impostazioni avanzate forniscono il controllo sulle opzioni che non appartengono specificamente ad altri gruppi di impostazioni.

Insieme di prevenzione del sovradattamento. Il metodo della rete neurale consente di separare internamente i record in un insieme di creazione del modello e un insieme di prevenzione del sovradattamento, che è un insieme indipendente di record di dati utilizzato per tenere traccia degli errori durante l'addestramento per impedire la modellazione della variazione casuale nei dati. Specificare una percentuale di record. Il valore predefinito è 30.

Replica risultati. L'impostazione di un seme random consente di replicare le analisi. Specificare un intero o fare clic su **Genera** per creare un intero pseudocasuale compreso tra 1 e 2147483647 incluso. Per default, le analisi vengono replicate con il seme 229176228.

Valori mancanti nei predittori. Specifica in che modo vengono trattati i valori mancanti. **Elimina listwise** rimuove dalla creazione del modello i record con valori mancanti nei predittori. **Assegna valori mancanti** sostituisce i valori mancanti nei predittori e utilizza tali record nell'analisi. I campi continui assegnano la media dei valori minimi e massimi osservati, i campi categoriali assegnano la categoria che ricorre più di frequente. Si noti che i record con valori mancanti relativi ad altri campi specificati nella scheda Campi vengono sempre eliminati dalla creazione del modello.

Opzioni del modello

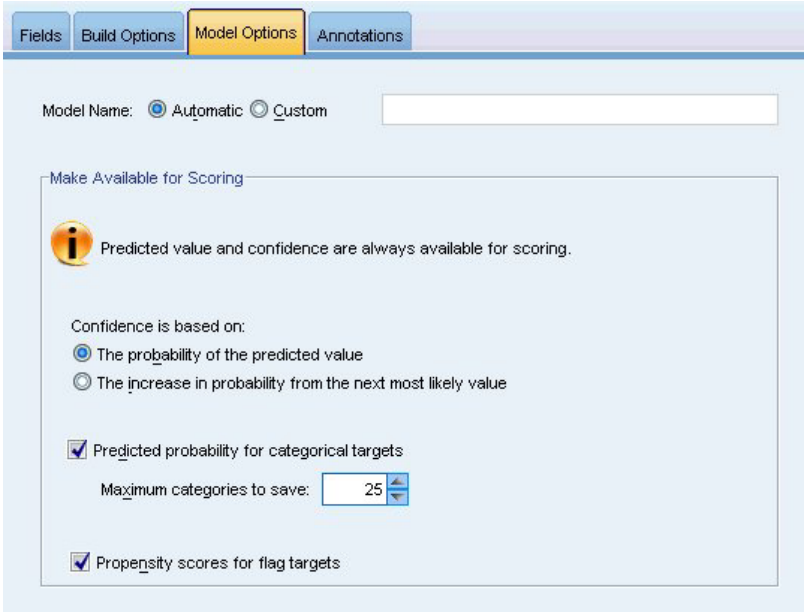


Figura 37. Scheda Opzioni modello

Nome modello. È possibile generare il nome del modello automaticamente in base ai campi obiettivo oppure specificare un nome personalizzato. Il nome generato automaticamente è il nome del campo obiettivo. Se sono presenti più obiettivi, il nome del modello è l'elenco dei nomi dei campi in ordine, collegati dalla "e" commerciale. Ad esempio se *campo1 campo2 campo3* sono obiettivi, il nome del modello è: *campo1 & campo2 & campo3*.

Rendi disponibile per il calcolo del punteggio. Quando viene calcolato il punteggio del modello, devono essere generati gli elementi selezionati in questo gruppo. Il valore previsto (per tutti gli obiettivi) e la confidenza (per gli obiettivi categoriali) vengono sempre calcolati durante il calcolo del punteggio del modello. La confidenza calcolata può essere basata sulla probabilità del valore atteso (la probabilità prevista più alta) o sulla differenza tra la probabilità prevista più alta e la seconda probabilità prevista più alta.

- **Probabilità prevista per gli obiettivi categoriali.** Genera le probabilità previste per gli obiettivi categoriali. Viene creato un campo per ogni categoria.
- **Punteggi di propensione per gli obiettivi flag.** Per i modelli con obiettivi di tipo flag (che restituiscono un risultato sì o nessuna previsione), è possibile richiedere i punteggi di propensione grezza che indicano la verosimiglianza del risultato vero specificato per il campo obiettivo. Il modello produce punteggi di propensione grezza; se le partizioni sono attive, il modello produce anche punteggi di propensione regolata basati sulla partizione di test.

Riepilogo del modello

Model Summary

Target	Previously defaulted
Model	Multilayer Perceptron
Stopping Rule Used	Error cannot be further decreased
Hidden Layer 1 Neurons	4

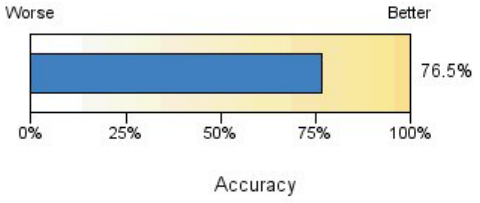


Figura 38. Vista Riepilogo del modello di reti neurali

La visualizzazione Riepilogo del modello è una snapshot della precisione della previsione o classificazione della rete neurale.

Riepilogo del modello. La tabella identifica l'obiettivo, il tipo di rete neurale addestrata, la regola di arresto che ha arrestato l'addestramento (visualizzata se è stata addestrata una rete percettrone multistrato) ed il numero di neuroni in ciascun livello nascosto della rete.

Qualità rete neurale. Il grafico visualizza la precisione del modello finale, visualizzato in formato ingrandito. Per un obiettivo categoriale, si tratta semplicemente della percentuale di record per i quali il valore previsto corrisponde al valore osservato. Per un target continuo, è 1 meno il rapporto tra l'errore assoluto della media nella previsione (la media dei valori assoluti dei valori previsti meno i valori osservati) e l'intervallo di valori previsti (il massimo valore previsto meno il minimo valore previsto).

Obiettivi multipli. Se sono presenti più obiettivi, ciascuno di essi è visualizzato nella riga **Obiettivo** della tabella. La precisione visualizzata nel grafico è la media delle singole precisioni degli obiettivi.

Importanza predittore

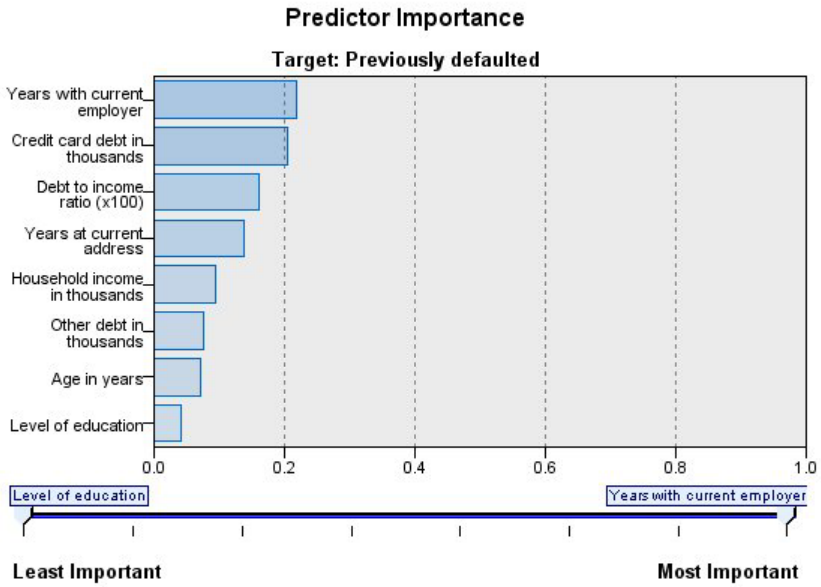


Figura 39. Visualizzazione Importanza predittore

Generalmente, lo sforzo della modellazione viene concentrato sui campi predittore più importanti, senza considerare o ignorando i campi di minore importanza. Il grafico dell'importanza dei predittori rende più semplice questa operazione, indicando l'importanza relativa di ciascun predittore nella stima del modello. Poiché i valori sono relativi, la somma dei valori visualizzata per tutti i predittori è 1.0. L'importanza dei predittori non è correlata alla precisione del modello. Riguarda unicamente l'importanza di ciascun predittore nell'esecuzione di una previsione e non il grado di precisione della previsione.

Obiettivi multipli. Se sono presenti più obiettivi, ciascun obiettivo viene visualizzato in un grafico separato ed è disponibile un elenco a discesa **Obiettivo** che controlla gli obiettivi da visualizzare.

Previsioni e osservazioni

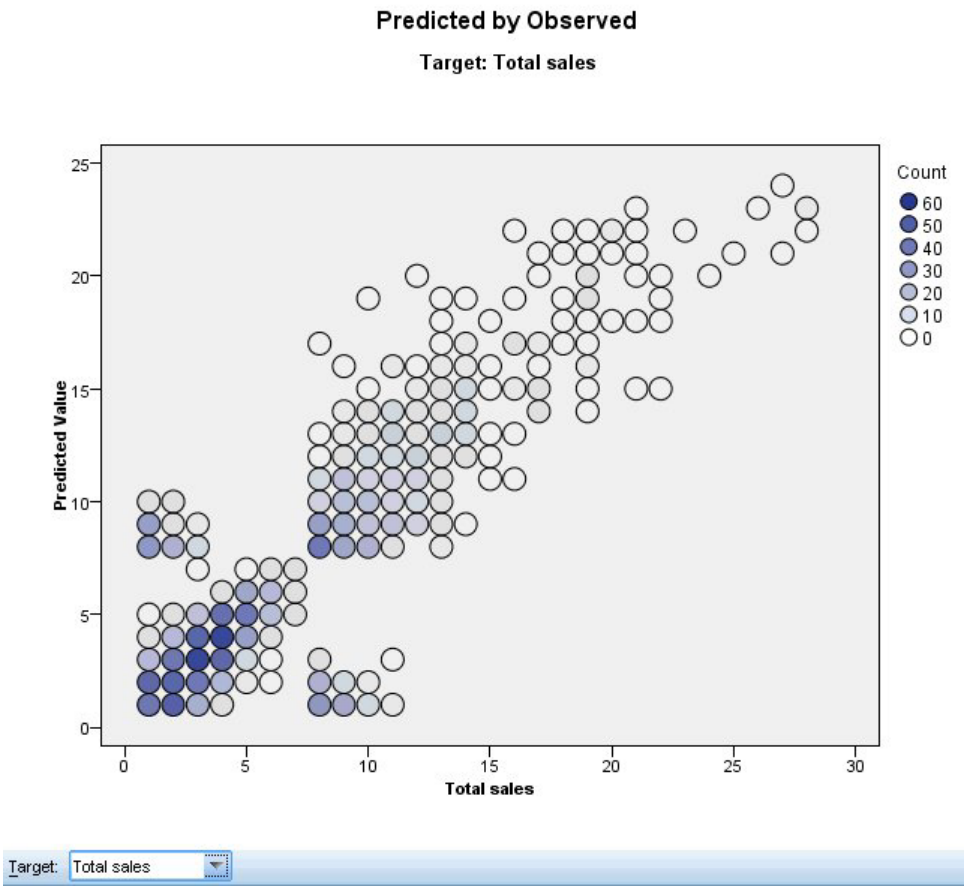


Figura 40. Visualizzazione Previsioni e osservazioni

Per i target continui, viene visualizzato un grafico a dispersione in bin dei valori previsti sull'asse verticale in base ai valori osservati sull'asse orizzontale.

Obiettivi multipli. Se sono presenti più target continui, ciascun target viene visualizzato in un grafico separato ed è disponibile un elenco a discesa **Target** che controlla il target da visualizzare.

Classificazione

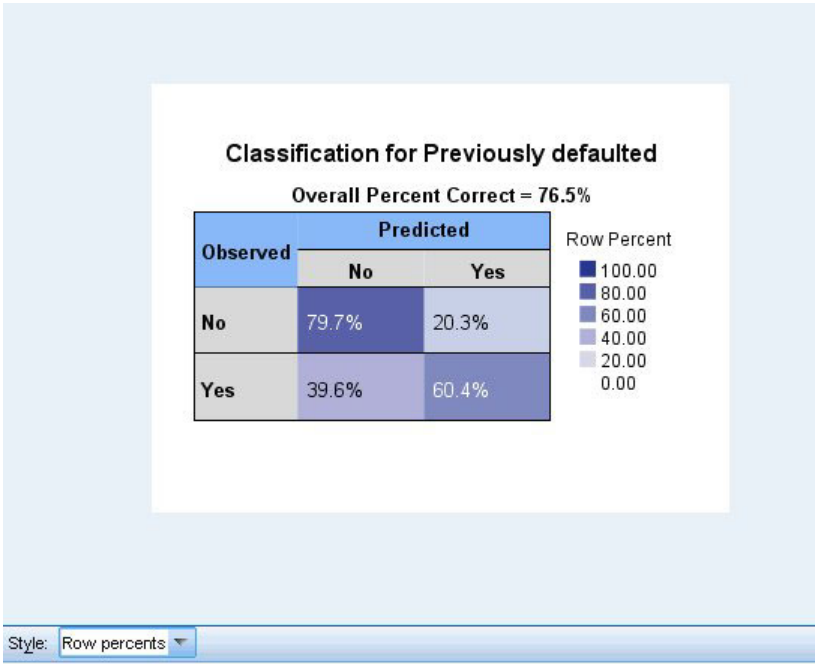


Figura 41. Visualizzazione Classificazione, stile percentuali di riga

Per gli obiettivi categoriali, visualizza la classificazione incrociata dei valori osservati e previsti in una mappa termica, più la percentuale globale corretta.

Stili di tabella. Sono disponibili diversi stili di visualizzazione, a cui è possibile accedere dall'elenco a discesa **Stile**.

- **Percentuali di riga.** Questa opzione visualizza la percentuale di righe (il numero di celle espresso come percentuale del totale di righe) nelle celle. Questa è l'opzione di default.
- **Conteggi delle celle.** Questa opzione visualizza il numero di celle nelle celle. L'ombreggiatura per la mappa termica è comunque basata sulle percentuali di riga.
- **Mappa termica.** Questa opzione non visualizza alcun valore nelle celle, ma solo l'ombreggiatura.
- **Compresso.** Questa opzione non visualizza alcuna intestazione di righe o colonne o valori nelle celle. Può essere utile se l'obiettivo ha molte categorie.

Mancante. Se, per alcuni record, non sono presenti valori nell'obiettivo, i valori vengono visualizzati in una riga **(Mancante)** al di sotto di tutte le righe valide. I record con valori mancanti non contribuiscono alla percentuale globale corretta.

Obiettivi multipli. Se sono presenti più obiettivi di categoria, ciascun obiettivo è visualizzato in una tabella separata ed è presente un elenco a discesa **Obiettivo** che controlla l'obiettivo da visualizzare.

Tabelle di grandi dimensioni. Se l'obiettivo visualizzato contiene più di 100 categorie, non viene visualizzata alcuna tabella.

Rete

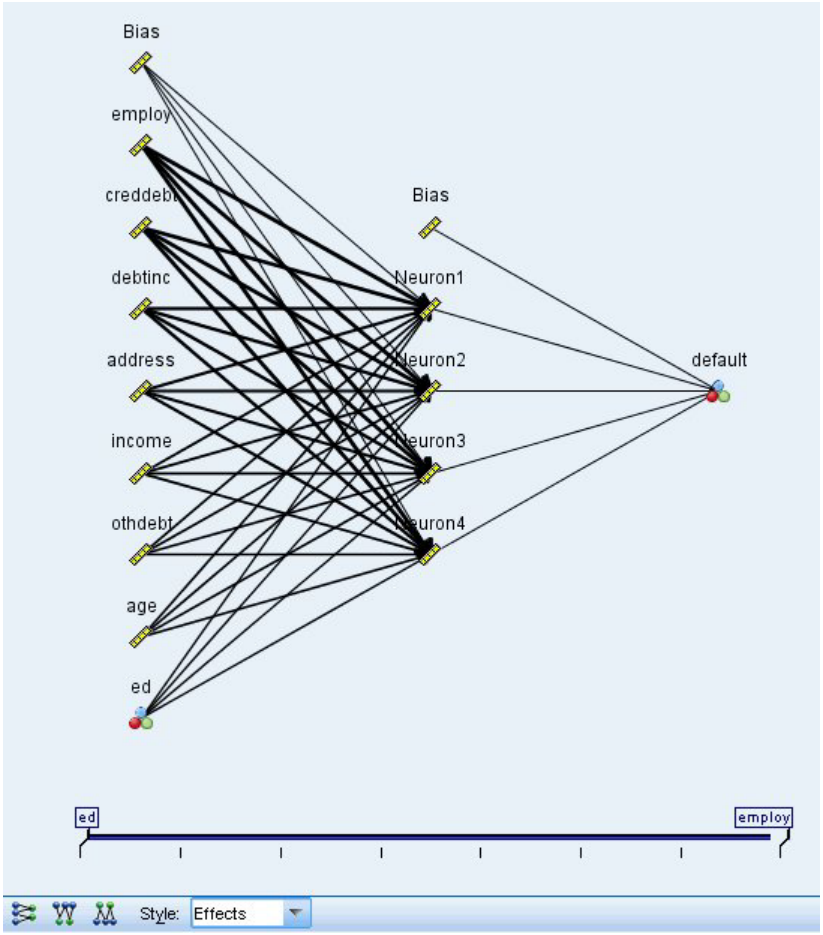


Figura 42. Visualizzazione Rete, input a sinistra, stile effetti

Viene visualizzata una rappresentazione grafica della rete neurale.

Stili del grafico. Sono disponibili due stili di visualizzazione differenti, a cui è possibile accedere dall'elenco a discesa **Stile**.

- **Effetti.** Visualizza ciascun predittore e obiettivo come un nodo unico nel diagramma, indipendentemente dal fatto che la scala di misurazione sia continua o relativa alla categoria. Questa è l'opzione di default.
- **Coefficienti.** Visualizza più nodi indicatori per gli obiettivi ed i predittori relativi alla categoria. Le linee di collegamento nel diagramma dei coefficienti sono colorate in base al valore stimato del peso sinaptico.

Orientamento del diagramma. Per default, nel diagramma di rete gli input sono visualizzati a sinistra e gli obiettivi a destra. Utilizzando i controlli della barra degli strumenti, è possibile modificare l'orientamento in modo tale da posizionare gli input in alto e gli obiettivi in basso oppure gli input in basso e gli obiettivi in alto.

Importanza predittore. Le linee di collegamento nel diagramma sono pesate in base all'importanza del predittore; le linee più larghe corrispondono ad un'importanza maggiore. Il dispositivo di scorrimento Importanza predittore nella barra degli strumenti consente di controllare i predittori visualizzati nel diagramma della rete. L'azione del dispositivo di scorrimento non modifica il modello, consente semplicemente di evidenziare i predittori più importanti.

Obiettivi multipli. Se sono presenti più obiettivi, sono tutti visualizzati nel grafico.

Impostazioni

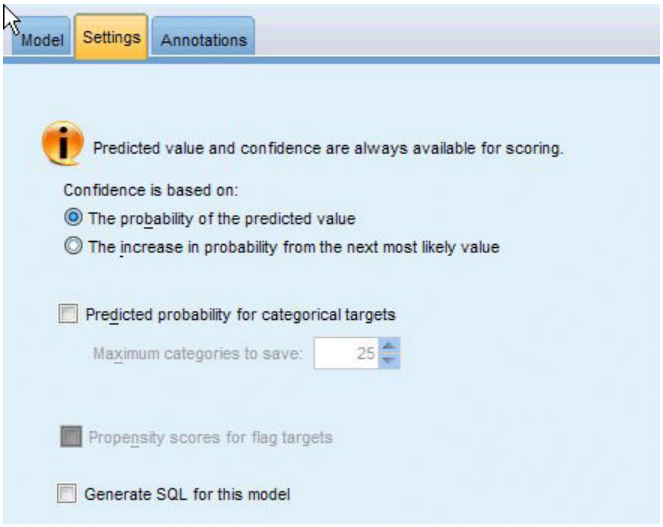


Figura 43. Scheda Impostazioni

Quando si esegue il calcolo del punteggio del modello, devono essere generati gli elementi selezionati in questa scheda. Il valore previsto (per tutti gli obiettivi) e la confidenza (per gli obiettivi categoriali) vengono sempre calcolati durante il calcolo del punteggio del modello. La confidenza calcolata può essere basata sulla probabilità del valore atteso (la probabilità prevista più alta) o sulla differenza tra la probabilità prevista più alta e la seconda probabilità prevista più alta.

- **Probabilità prevista per gli obiettivi categoriali.** Genera le probabilità previste per gli obiettivi categoriali. Viene creato un campo per ogni categoria.
- **Punteggi di propensione per gli obiettivi flag.** Per i modelli con obiettivi di tipo flag (che restituiscono un risultato sì o nessuna previsione), è possibile richiedere i punteggi di propensione grezza che indicano la verosimiglianza del risultato vero specificato per il campo obiettivo. Il modello produce punteggi di propensione grezza; se le partizioni sono attive, il modello produce anche punteggi di propensione regolata basati sulla partizione di test.

Genera SQL per questo modello. Quando si utilizzano dati da un database, è possibile "restituire" codice SQL al database per l'esecuzione, migliorando in tal modo le performance di molte operazioni.

Calcola il punteggio convertendo in SQL nativo. Se questa opzione è selezionata, genera il codice SQL per calcolare in modo nativo il punteggio del modello all'interno dell'applicazione.

Capitolo 9. Elenco di decisioni

I modelli Elenco di decisioni identificano i sottogruppi o i **segmenti** che mostrano una probabilità maggiore o minore che si verifichi un determinato risultato binario (sì o no) rispetto al campione globale. Per esempio, è possibile che si cerchino i clienti a minore rischio di abbandono o quelli che più probabilmente rispondano in modo favorevole a una determinato offerta o campagna. Il Visualizzatore dell'elenco di decisioni fornisce controllo completo sul modello, consentendo di modificare segmenti, aggiungere le proprie regole di business, specificare il modo in cui viene calcolato il punteggio di ciascun segmento e personalizzare il modello in diversi modi per ottimizzare la proporzione di risultati tra tutti i segmenti. È pertanto particolarmente indicato per generare mailing list o per identificare i record da utilizzare come obiettivo per una determinata campagna. È anche possibile utilizzare più **attività di mining** per combinare gli approcci di modellazione — ad esempio, identificando segmenti con prestazioni elevate o ridotte all'interno dello stesso modello ed includendo oppure escludendo ciascuno di essi in modo appropriato nella fase di calcolo del punteggio.

Segmenti, regole e condizioni

Un modello consiste in un elenco di segmenti, ognuno dei quali viene definito da una regola che seleziona i record corrispondenti. Una determinata regola può avere più condizioni, per esempio:

```
RFM_SCORE > 10 and
MONTHS_CURRENT <= 9
```

Le regole vengono applicate nell'ordine elencato, con la prima regola corrispondente che determina il risultato per un dato record. Prese indipendentemente, le regole o le condizioni possono sovrapporsi, ma l'ordine delle regole risolve eventuali ambiguità. Se nessuna regola corrisponde, il record viene assegnato alla regola Resto.

Controllo completo sul calcolo del punteggio

Il Visualizzatore dell'elenco di decisioni consente di visualizzare, modificare e riorganizzare i segmenti e di scegliere i segmenti da includere o escludere per il calcolo del punteggio. Per esempio, è possibile scegliere di escludere un gruppo di clienti dalle offerte future e includerne altri e vedere immediatamente come questa scelta influisce sul tasso di risultati complessivo. I modelli Elenco di decisioni restituiscono il punteggio *Sì* per i segmenti inclusi e *$null$* per tutti gli altri elementi, incluso il resto. Questo controllo diretto del calcolo del punteggio rende i modelli Elenco di decisioni particolarmente indicati per la creazione di mailing list; questi modelli vengono ampiamente utilizzati nelle operazioni di CRM (Customer Relationship Management), comprese le applicazioni di call center o marketing.

Attività di mining, misure e selezioni

Il processo di creazione dei modelli è condotto tramite **attività di mining**. Ogni attività di mining avvia effettivamente un nuovo processo di modellazione e restituisce un nuovo insieme di modelli alternativi tra cui scegliere. L'attività di default è basata sulle specifiche iniziali del nodo Elenco di decisioni, ma è possibile definire il numero desiderato di attività personalizzate. È anche possibile applicare le attività in modo iterativo — ad esempio, è possibile eseguire una ricerca di probabilità sull'intero insieme di addestramento ed eseguire una ricerca probabilità bassa sul resto per eliminare i segmenti con prestazioni ridotte.

Selezioni di dati

È possibile definire selezioni di dati e misure di modelli personalizzate per la creazione e la valutazione del modello. Per esempio, è possibile specificare una selezione di dati in un'attività di mining per adattare il modello a una regione specifica e creare una misura personalizzata per valutare la

performance del modello a livello nazionale. A differenza delle attività di mining, le misure non modificano il modello sottostante, ma forniscono un ulteriore strumento per valutarne la performance.

Aggiunta delle conoscenze di business

Ottimizzando oppure estendendo i segmenti identificati dall'algoritmo, Visualizzatore dell'elenco di decisioni consente di incorporare le proprie conoscenze di business nel modello. È possibile modificare i segmenti generati dal modello o aggiungere ulteriori segmenti in base alle regole specificate. È quindi possibile applicare le modifiche e visualizzare in anteprima i risultati.

Per un ulteriore approfondimento, un collegamento dinamico consente di esportare i dati in Excel, dove i dati possono essere utilizzati per creare grafici di presentazione e calcolare misure personalizzate, come ROI e profitti complessi, che possono essere visualizzate nel Visualizzatore dell'elenco di decisioni durante la creazione del modello.

Esempio. La divisione di marketing di una società finanziaria desidera ottenere risultati più redditizi nelle campagne future, inviando offerte mirate ai singoli clienti. È possibile utilizzare un modello Elenco di decisioni per identificare le caratteristiche dei clienti che hanno più probabilità di rispondere in modo favorevole in base alle promozioni precedenti e generare una mailing list in base ai risultati.

Requisiti. Un unico campo obiettivo categoriale con livello di misurazione di tipo *Flag* o *Nominale* che indica il risultato binario che si desidera prevedere (sì/no) e almeno un campo di input. Quando il campo obiettivo è di tipo *Nominale*, è necessario scegliere manualmente un unico valore da trattare come **risultato** o **risposta**; tutti gli altri valori vengono raggruppati come **non risultati**. È inoltre possibile specificare un campo frequenza facoltativo. I campi data/ora continui vengono ignorati. Gli input di intervalli numerici continui vengono automaticamente discretizzati dall'algoritmo, come specificato nella scheda Livello avanzato del nodo Modelli. Per un controllo più accurato della discretizzazione, aggiungere un nodo Discretizza a monte e utilizzare il campo nei bin come input con livello di misurazione *Ordinale*.

Opzioni del modello Elenco di decisioni

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.

Utilizza dati partizionati. Se è definito un campo partizione, questa opzione garantisce che per la creazione del modello vengano utilizzati dati solo dalla partizione di addestramento.

Crea modelli di suddivisione. Crea un modello separato per ogni valore possibile dei campi di input che sono stati specificati come campi di suddivisione. Per ulteriori informazioni, consultare l'argomento “Creazione di modelli di suddivisione” a pagina 28.

Modalità. Specifica il metodo utilizzato per la creazione del modello.

- **Genera modello.** Genera automaticamente un modello sulla palette dei modelli quando il nodo viene eseguito. Il modello risultante può essere aggiunto ai flussi per il calcolo del punteggio ma non può essere ulteriormente modificato.
- **Avvia sessione interattiva.** Visualizza la finestra di modellazione di Visualizzatore dell'elenco di decisioni interattiva, che consente di scegliere tra diverse alternative e di applicare ripetutamente l'algoritmo con impostazioni differenti per modificare o espandere il modello in modo progressivo. Per ulteriori informazioni, consultare l'argomento “Visualizzatore dell'elenco di decisioni” a pagina 145.
- **Utilizza informazioni sulla sessione interattiva salvata.** Avvia una sessione interattiva utilizzando le impostazioni salvate in precedenza. Le impostazioni interattive possono essere salvate dal Visualizzatore dell'elenco di decisioni mediante il menu Genera (per creare un modello o un nodo Modelli) o mediante il menu File (per aggiornare il nodo da cui la sessione è stata avviata).

Valore obiettivo. Specifica il valore del campo obiettivo indicante il risultato che si desidera modellare. Per esempio, se il tasso di abbandono del campo obiettivo è codificato come `0 = no` e `1 = sì`, specificare `1` per individuare le regole che indicano per quali record esistono probabilità di abbandono.

Trova segmenti con. Indica se la ricerca per la variabile obiettivo deve eseguire la ricerca di una **Probabilità elevata** o di una **Probabilità bassa** di occorrenza. L'individuazione e l'esclusione di tali segmenti possono costituire un valido sistema per migliorare il modello e possono rivelarsi particolarmente utili quando i segmenti residui hanno una probabilità bassa.

Numero massimo di segmenti. Specifica il numero massimo di segmenti da restituire. Vengono creati i primi *N* segmenti in cui il segmento migliore è quello con la massima probabilità oppure, se più modelli hanno le stesse probabilità, quello con la massima copertura. L'impostazione minima consentita è 1; non è prevista un'impostazione massima.

Dimensione minima del segmento. Le due impostazioni riportate sotto determinano la dimensione minima del segmento. Il valore maggiore tra i due ha la precedenza. Per esempio, se il valore della percentuale corrisponde a un numero maggiore del valore assoluto, l'impostazione della percentuale avrà la precedenza.

- **Come percentuale del segmento precedente (%).** Specifica le dimensioni minime del gruppo come una percentuale di record. L'impostazione minima consentita è 0; l'impostazione massima consentita è 99.9.
- **Come valore assoluto (N).** Specifica le dimensioni minime del gruppo come un numero assoluto di record. L'impostazione minima consentita è 1; non è prevista un'impostazione massima.

Regole di segmento.

Numero massimo di attributi. Specifica il numero massimo di condizioni per ogni regola di segmento. L'impostazione minima consentita è 1; non è prevista un'impostazione massima.

- **Consenti riutilizzo attributo.** Quando questa opzione è attivata, ogni ciclo può considerare tutti gli attributi, persino quelli già utilizzati nei cicli precedenti. Le condizioni relative a un segmento vengono create in cicli: ogni ciclo aggiunge una nuova condizione. Il numero dei cicli viene definito mediante l'impostazione **Numero massimo di attributi**.

Intervallo di confidenza per le nuove condizioni (%). Specifica il livello di confidenza per la verifica della significatività di un segmento. Questa impostazione influisce in maniera significativa sul numero degli eventuali segmenti restituiti, nonché sul numero di condizioni per regola di segmento. Più alto è il valore, più piccolo sarà l'insieme di risultati restituiti. L'impostazione minima consentita è 50; l'impostazione massima consentita è 99.9.

Opzioni avanzate del nodo Elenco di decisioni

Le opzioni avanzate consentono di definire con precisione il processo di creazione dei modelli.

Metodo di raccolta. Il metodo utilizzato per la discretizzazione dei campi continui (conteggio uguale o larghezza uguale).

Numero di bin. Il numero di bin da creare per i campi continui. L'impostazione minima consentita è 2; non è prevista un'impostazione massima.

Larghezza di ricerca del modello. Il numero massimo di risultati del modello per ciclo utilizzabile per il ciclo successivo. L'impostazione minima consentita è 1; non è prevista un'impostazione massima.

Larghezza di ricerca della regola. Il numero massimo di risultati di regola per ciclo utilizzabile per il ciclo successivo. L'impostazione minima consentita è 1; non è prevista un'impostazione massima.

Fattore di unione bin. La quantità minima di cui un segmento deve aumentare quando viene unito all'elemento adiacente. L'impostazione minima consentita è 1.01; non è prevista un'impostazione massima.

- **Consenti valori mancanti nelle condizioni.** `Vero` per consentire la verifica `IS MISSING` nelle regole.
- **Scarta risultati intermedi.** Se `Vero`, vengono restituiti solo i risultati finali della ricerca. Un risultato finale è un risultato che non viene ulteriormente rifinito nel processo di ricerca. Se `Falso`, vengono restituiti anche i risultati intermedi.

  Numero massimo di alternative. Specifica il numero massimo di alternative che possono essere restituite con l'esecuzione dell'attività di mining. L'impostazione minima consentita è 1; non è prevista un'impostazione massima.

  L'attività di mining restituisce solo il numero effettivo di alternative, fino al numero massimo specificato. Per esempio, se il massimo è impostato su 100 e vengono trovate solo 3 alternative, verranno visualizzate solo quelle 3.

Nugget del modello Elenco di decisioni

Un modello consiste in un elenco di **segmenti**, ognuno dei quali viene definito da una **regola** che seleziona i record corrispondenti. È possibile visualizzare o modificare con facilità i segmenti prima di generare il modello e scegliere quelli che si desidera includere o escludere. Quando vengono utilizzati nel calcolo del punteggio, i modelli Elenco di decisioni restituiscono *Sì* per i segmenti inclusi e *$null$* in tutti gli altri casi, compreso il resto. Questo controllo diretto del calcolo del punteggio rende i modelli Elenco di decisioni particolarmente indicati per la creazione di mailing list; questi modelli vengono ampiamente utilizzati nelle operazioni di CRM (Customer Relationship Management), comprese le applicazioni di call center o marketing.

Quando si esegue un flusso contenente un modello Elenco di decisioni, il nodo aggiunge tre nuovi campi contenenti il punteggio, *1* (vale a dire *Sì*) per i campi inclusi o *$null* per i campi esclusi, la probabilità (la percentuale di risultati) relativa al segmento nel quale rientra il record e il numero ID del segmento. I nomi dei nuovi campi derivano dal nome del campo di output di cui si sta eseguendo la previsione, a cui viene aggiunto il prefisso *$D-* per il punteggio, *$DP-* per la probabilità e *$DI-* per l'ID del segmento.

Il punteggio del modello viene calcolato in base al valore obiettivo specificato quando il modello è stato creato. È possibile escludere i segmenti manualmente, in modo che il relativo punteggio risulti *$null$*. Ad esempio, se viene eseguita una ricerca di bassa probabilità per individuare segmenti con un numero di risultati inferiore alla media, per tali segmenti "con valore basso" il punteggio verrà calcolato come *Sì* a meno che non vengano esclusi manualmente. Se necessario, i valori null possono essere ricodificati come *No* mediante un nodo Ricava o un nodo Riempimento.

PMML

Un modello Elenco di decisioni può essere archiviato come RuleSetModel PMML con un criterio di selezione "primo risultato". Tuttavia, tutte le regole devono avere lo stesso punteggio. Per tenere conto delle modifiche apportate al campo obiettivo o al valore obiettivo, più modelli Insieme di regole possono essere archiviati in un unico file e applicati in ordine: i casi per cui non si trovano corrispondenze nel primo modello vengono passati al secondo e così via. Il nome dell'algoritmo *DecisionList* viene utilizzato per indicare questo comportamento non standard e solo i modelli Insieme di regole denominati in questo modo vengono riconosciuti (e i relativi punteggi calcolati) come modelli Elenco di decisioni.

Impostazioni del nugget del modello Elenco di decisioni

La scheda Impostazioni di un nugget del modello Elenco di decisioni consente di ottenere i punteggi di propensione e di abilitare o disabilitare l'ottimizzazione SQL. Questa scheda è disponibile solo dopo che il nugget del modello è stato aggiunto a un flusso.

Calcola punteggi di propensione grezza. Per i modelli con obiettivi di tipo flag (che restituiscono un risultato sì o nessuna previsione), è possibile richiedere i punteggi di propensione grezza che indicano la

verosimiglianza del risultato vero specificato per il campo obiettivo. Tali punteggi si aggiungono ai valori di previsione e confidenza che potrebbero venire generati durante il calcolo del punteggio.

Calcola punteggi di propensione regolata. I punteggi di propensione grezza sono basati esclusivamente sui dati di addestramento e potrebbero essere eccessivamente ottimistici a causa della tendenza di molti modelli a sovradattare questi dati. Le propensioni regolate cercano di operare una compensazione esaminando le prestazioni del modello rispetto a una partizione di test o di convalida. Per utilizzare questa opzione è necessario che nel flusso sia definito un campo partizione e che nel nodo Modelli siano attivati i punteggi di propensione regolati, prima di generare il modello.

Calcola il punteggio convertendo in SQL nativo. Se questa opzione è selezionata, genera il codice SQL per calcolare in modo nativo il punteggio del modello all'interno dell'applicazione.

Visualizzatore dell'elenco di decisioni

L'interfaccia grafica di Visualizzatore dell'elenco di decisioni, basata sulle attività e di facile utilizzo, rende più semplice il processo di creazione del modello perché evita di doversi occupare dei dettagli elementari delle tecniche di data mining e consente di concentrarsi sulle parti dell'analisi che richiedono l'intervento da parte dell'utente, come la definizione degli obiettivi, la scelta dei gruppi obiettivo, l'analisi dei risultati e la selezione del modello ottimale.

Riquadro Modello di lavoro

Il riquadro Modello di lavoro visualizza il modello corrente insieme alle attività di mining e alle altre azioni valide per il modello di lavoro.

ID. Identifica l'ordine sequenziale dei segmenti. I segmenti di modello vengono calcolati in sequenza in base ai rispettivi ID.

Regole di segmento. Visualizza il nome del segmento e le condizioni definite per il segmento. Per default, il nome del segmento è rappresentato dal nome del campo o dai nomi dei campi concatenati utilizzati nelle condizioni, separati da una virgola.

Punteggio. Rappresenta il campo per cui si desidera eseguire la previsione, il cui valore si presuppone correlato ai valori di altri campi (i predittori).

Nota: la visualizzazione delle opzioni riportate di seguito può essere attivata o disattivata tramite la finestra di dialogo “Organizzazione delle misure del modello” a pagina 155.

Copertura. Il grafico a torta rappresenta visivamente la copertura di ogni segmento rispetto alla copertura complessiva.

Copertura (n). Mostra la copertura di ogni segmento rispetto alla copertura complessiva.

Frequenza. Elenca il numero di risultati ottenuti in relazione alla copertura. Per esempio, quando la copertura è 79 e la frequenza 50, per il segmento selezionato la risposta è stata di 50 su 79.

Probabilità. Indica la probabilità del segmento. Per esempio, quando la copertura è 79 e la frequenza 50, per quel segmento la probabilità è del 63.29% (50 diviso 79).

Errore. Indica l'errore del segmento.

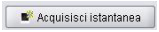
Le informazioni nella parte inferiore del riquadro indicano la copertura, la frequenza e la probabilità per tutto il modello.

Barra degli strumenti del modello di lavoro

Il riquadro Modello di lavoro consente di accedere alle seguenti funzioni tramite una barra degli strumenti.

Nota: alcune funzioni sono disponibili anche facendo clic con il tasto destro del mouse su un segmento di modello.

Tabella 9. Pulsanti della barra degli strumenti del modello di lavoro.

Pulsante della barra degli strumenti	**Descrizione**
	Visualizza la finestra di dialogo Genera nuovo modello, che contiene le opzioni per la creazione di un nuovo nugget del modello.
	Salva lo stato corrente della sessione interattiva. Il nodo Modelli Elenco di decisioni viene aggiornato con le impostazioni correnti, incluse le attività di mining, le snapshot dei modelli, le selezioni di dati e le misure personalizzate. Per riportare la sessione a questo stato, selezionare la casella di controllo **Utilizza informazioni sulla sessione interattiva salvata** nella scheda Modello del nodo Modelli e fare clic su **Esegui**.
	Visualizza la finestra di dialogo Organizza misure del modello. Per ulteriori informazioni, consultare l'argomento “Organizzazione delle misure del modello” a pagina 155.
	Visualizza la finestra di dialogo Organizza selezioni di dati. Per ulteriori informazioni, consultare l'argomento “Organizzazione delle selezioni di dati” a pagina 151.
	Visualizza la scheda Snapshot. Per ulteriori informazioni, consultare l'argomento “Scheda Snapshot” a pagina 147.
	Visualizza la scheda Alternative. Per ulteriori informazioni, consultare l'argomento “Scheda Alternative”.
	Scatta una snapshot della struttura del modello corrente. Le snapshot vengono visualizzate nella scheda Snapshot e sono utilizzate generalmente per eseguire confronti tra i modelli.
	Visualizza la finestra di dialogo Inserimento di segmenti, che fornisce le opzioni per la creazione di nuovi segmenti del modello.
	Visualizza la finestra di dialogo Modifica delle regole di segmento, che contiene le regole per l'aggiunta delle condizioni ai segmenti del modello o la modifica delle condizioni del segmento del modello precedentemente definite.
	Sposta il segmento selezionato verso l'alto nella gerarchia del modello.
	Sposta il segmento selezionato verso il basso nella gerarchia del modello.
	Elimina il segmento selezionato.
	Attiva/disattiva l'inclusione del segmento selezionato nel modello. Quando il segmento viene escluso, i relativi risultati vengono aggiunti al resto. Questa operazione è diversa dall'eliminazione di un segmento poiché il segmento può essere riattivato.

Scheda Alternative

La scheda Alternative, generata quando si fa clic su **Trova segmenti**, elenca tutti i risultati alternativi del mining per il modello o il segmento selezionato nel riquadro Modello di lavoro.

Per utilizzare un'alternativa come modello di lavoro, evidenziare l'alternativa desiderata e fare clic su **Carica**; il modello alternativo viene visualizzato nel riquadro Modello di lavoro.

Nota: la scheda Alternative viene visualizzata solo se è stata impostata l'opzione **Numero massimo di alternative** nella scheda Livello avanzato del nodo di modellazione Elenco di decisioni per creare più di una alternativa.

Ogni alternativa di modello generata visualizza informazioni specifiche sul modello:

Nome. Ogni alternativa è contrassegnata da un numero sequenziale. La prima alternativa in genere è quella che contiene i risultati migliori.

Obiettivo. Indica il valore obiettivo. Ad esempi: 1, che corrisponde a "true".

N. di segmenti. Il numero di regole di segmento utilizzate nel modello alternativo.

Copertura. La copertura del modello alternativo.

Freq. Il numero di risultati in relazione alla copertura.

Prob. Indica la percentuale di probabilità del modello alternativo.

Nota: i risultati alternativi non vengono salvati insieme al modello; i risultati sono validi solo durante la sessione attiva.

Scheda Snapshot

Una snapshot è la visualizzazione di un modello in un determinato momento. Per esempio, si può acquisire la snapshot di un modello quando si desidera caricare un modello alternativo diverso nel riquadro Modello di lavoro senza perdere il lavoro effettuato sul modello corrente. La scheda Snapshot elenca tutte le snapshot dei modelli acquisite manualmente per qualsiasi numero di stati del modello di lavoro.

Nota: le snapshot vengono salvate insieme al modello. Si consiglia di eseguire una snapshot quando si carica il primo modello. Questa snapshot manterrà la struttura del modello originale consentendo di tornare sempre allo stato originario del modello. Il nome della snapshot generata viene visualizzato sotto forma di timestamp per indicare quando è stata generata.

Creazione di una snapshot di un modello

1. Selezionare un modello o un'alternativa appropriata da visualizzare nel riquadro Modello di lavoro.
2. Apportare le eventuali modifiche necessarie al modello di lavoro.
3. Fare clic su **Acquisisci snapshot**. Nella scheda Snapshot viene visualizzata una nuova snapshot.
   Nome. Il nome della snapshot, Per modificare il nome di una snapshot, fare doppio clic sul nome.
   Obiettivo. Indica il valore obiettivo. Ad esempi: 1, che corrisponde a "true".
   N. di segmenti. Il numero di regole di segmento utilizzate nel modello.
   Copertura. La copertura del modello.
   Freq. Il numero di risultati in relazione alla copertura.
   Prob. Indica la percentuale di probabilità del modello.
4. Per utilizzare una snapshot come modello di lavoro, evidenziare la snapshot desiderata e fare clic su **Carica**; il modello della snapshot viene visualizzato nel riquadro Modello di lavoro.
5. Per eliminare una snapshot è possibile scegliere **Elimina** o fare clic con il pulsante destro del mouse sull'istantanea e scegliere **Elimina** dal menu.

Utilizzo di Visualizzatore dell'elenco di decisioni

La creazione di un modello in grado di prevedere nel modo migliore la risposta e il comportamento dei clienti avviene in varie fasi. All'avvio di Visualizzatore dell'elenco di decisioni, nel modello di lavoro

vengono inseriti automaticamente i segmenti di modello e le misure definiti e l'applicazione è così pronta per avviare un'attività di mining, modificare i segmenti o le misure in funzione delle esigenze e generare un nuovo modello o un nuovo nodo Modelli.

È possibile aggiungere una o più regole di segmento fino a sviluppare un modello soddisfacente. Le regole di segmento possono essere aggiunte al modello eseguendo attività di mining o utilizzando la funzione **Modifica regola segmenti**.

Durante la creazione del modello è possibile valutare le prestazioni del modello eseguendone la convalida con i dati delle misure, visualizzando il modello sotto forma di grafico o generando misure personalizzate in Excel.

Quando la qualità del modello risulta soddisfacente, è possibile generarne uno nuovo e collocarlo nell'area o nella palette Modelli di IBM SPSS Modeler.

Attività di mining

Un'**attività di mining** è una raccolta di parametri che determina il modo in cui vengono generate nuove regole. Alcuni di questi parametri sono selezionabili, per consentire all'utente di adattare in modo flessibile i modelli a situazioni nuove. Un'attività è costituita da un modello di attività (tipo), da un obiettivo e da una selezione per la creazione (insieme di dati di mining).

Le sezioni seguenti illustrano in modo dettagliato le varie operazioni relative alle attività di mining:

- "Esecuzione di attività di mining"
- "Creazione e modifica di un'attività di mining" a pagina 149
- "Organizzazione delle selezioni di dati" a pagina 151

Esecuzione di attività di mining: Visualizzatore dell'elenco di decisioni consente di aggiungere manualmente regole di segmento ad un modello eseguendo attività di mining oppure copiando ed incollando le regole di segmento tra i modelli. Un'attività di mining contiene informazioni su come generare nuove regole di segmento (le impostazioni dei parametri di data mining, quali la strategia di ricerca, gli attributi di origine, la larghezza di ricerca, il livello di confidenza e così via), sul comportamento dei clienti da prevedere e i dati da analizzare. Lo scopo di un'attività di mining è la ricerca delle migliori regole di segmento possibili.

Per generare una regola di segmento di modello con l'esecuzione di un'attività di mining:

1. Fare clic sulla riga **Resto**. Se nel riquadro del modello di lavoro sono già visualizzati dei segmenti è possibile anche selezionare uno dei segmenti per individuare altre regole in base al segmento selezionato. Dopo avere selezionato il resto o il segmento, generare il modello o i modelli alternativi utilizzando uno dei seguenti metodi:

- Scegliere **Trova segmenti** dal menu Strumenti.
- Fare clic con il pulsante destro del mouse sul segmento o sulla riga **Resto** e scegliere **Trova segmenti**.
- Fare clic sul pulsante **Trova segmenti** nel riquadro Modello di lavoro.

Durante l'elaborazione dell'attività, nella parte inferiore dello spazio di lavoro viene visualizzato lo stato di avanzamento e viene segnalato il completamento dell'attività. Il tempo esatto necessario per completare un'attività dipende dalla complessità dell'attività di mining e dalle dimensioni dell'insieme di dati. Se nei risultati c'è un unico modello, questo viene visualizzato nel riquadro Modello di lavoro non appena l'attività viene completata; tuttavia, se i risultati contengono più modelli, questi vengono visualizzati nella scheda Alternative.

Nota: il risultato di un'attività sarà completato con modelli, completato senza modelli o non riuscito.

Il processo di individuazione di nuove regole di segmento può essere ripetuto fino a quando al modello non viene più aggiunta alcuna regola. Ciò significa che tutti i gruppi di clienti significativi sono stati individuati.

Le attività di mining si possono eseguire su qualsiasi segmento di modello esistente. Se il risultato di un'attività non corrisponde alle aspettative, è possibile decidere di avviare un'altra attività di mining sullo stesso segmento. In questo modo saranno individuate altre regole in base al segmento selezionato. I segmenti che si trovano "al di sotto" del segmento selezionato (ovvero che sono stati aggiunti al modello dopo il segmento selezionato) vengono sostituiti dai nuovi segmenti perché ogni segmento dipende da quelli che lo precedono.

Creazione e modifica di un'attività di mining: Un'attività di mining è il meccanismo che cerca la raccolta di regole che costituiscono un modello di dati. Oltre ai criteri di ricerca definiti nel modello selezionato, un'attività definisce anche l'obiettivo (la domanda effettiva che ha motivato l'analisi, per esempio quanti clienti è probabile che rispondano a un mailing) e individua gli insiemi di dati da utilizzare. Lo scopo di un'attività di mining è la ricerca dei migliori modelli possibili.

Creazione di un'attività di mining

Per creare un'attività di mining:

1. Selezionare il segmento su cui si desidera effettuare il mining per individuare altre condizioni per il segmento.
2. Fare clic su **Impostazioni**. Viene visualizzata la finestra di dialogo Crea/Modifica attività di mining. Questa finestra contiene una serie di opzioni per definire l'attività di mining.
3. Apportare le modifiche desiderate e fare clic su **OK** per tornare al riquadro Modello di lavoro. Visualizzatore dell'elenco di decisioni utilizza le impostazioni come valori di default per eseguire le singole attività finché non vengono selezionate impostazioni o attività alternative.
4. Fare clic su **Trova segmenti** per avviare l'attività di mining sul segmento selezionato.

Modifica di un'attività di mining

La finestra di dialogo Crea/Modifica attività di mining contiene una serie di opzioni per definire una nuova attività di mining o per modificarne una esistente.

Quasi tutti i parametri disponibili per le attività di mining sono simili a quelli presentati nel nodo Elenco di decisioni. Di seguito sono indicate le eccezioni. Per ulteriori informazioni, consultare l'argomento “Opzioni del modello Elenco di decisioni” a pagina 142.

Carica impostazioni: quando è stata creata più di un'attività di mining, selezionare l'attività richiesta.

Nuovo... Fare clic per creare una nuova attività di mining in base alle impostazioni dell'attività attualmente visualizzata.

Obiettivo

Campo obiettivo: rappresenta il campo per cui si desidera eseguire la previsione, il cui valore si presuppone correlato ai valori di altri campi (i predittori).

Valore obiettivo. Specifica il valore del campo obiettivo indicante il risultato che si desidera modellare. Per esempio, se il tasso di abbandono del campo obiettivo è codificato come `0 = no` e `1 = sì`, specificare `1` per individuare le regole che indicano per quali record esistono probabilità di abbandono.

Impostazioni Livello base

Numero massimo di alternative. Specifica il numero di alternative che saranno visualizzate dopo l'esecuzione dell'attività di mining. L'impostazione minima consentita è 1; non è prevista un'impostazione massima.

Impostazioni Livello avanzato

Modifica... Visualizza la finestra di dialogo **Modifica parametri avanzati** che consente di definire le impostazioni avanzate. Per ulteriori informazioni, consultare l'argomento "Modifica parametri avanzati".

Dati

Selezione per creazione. Presenta una serie di opzioni che consentono di specificare la misura di valutazione che deve essere analizzata da Visualizzatore dell'elenco di decisioni per individuare nuove regole. Le misure di valutazione elencate vengono create/modificate nella finestra di dialogo Organizza le selezioni di dati.

Campi disponibili. Presenta una serie di opzioni che consentono di visualizzare tutti i campi o di selezionare manualmente i campi da visualizzare.

Modifica... Se l'opzione **Personalizza** è selezionata, visualizza la finestra di dialogo **Personalizza campi disponibili**, che consente di selezionare i campi disponibili come attributi di segmenti trovati dall'attività di mining. Per ulteriori informazioni, consultare l'argomento "Personalizza campi disponibili".

Modifica parametri avanzati: La finestra di dialogo Modifica parametri avanzati contiene le seguenti opzioni di configurazione.

Metodo di raccolta. Il metodo utilizzato per la discretizzazione dei campi continui (conteggio uguale o larghezza uguale).

Numero di bin. Il numero di bin da creare per i campi continui. L'impostazione minima consentita è 2; non è prevista un'impostazione massima.

Larghezza di ricerca del modello. Il numero massimo di risultati del modello per ciclo utilizzabile per il ciclo successivo. L'impostazione minima consentita è 1; non è prevista un'impostazione massima.

Larghezza di ricerca della regola. Il numero massimo di risultati di regola per ciclo utilizzabile per il ciclo successivo. L'impostazione minima consentita è 1; non è prevista un'impostazione massima.

Fattore di unione bin. La quantità minima di cui un segmento deve aumentare quando viene unito all'elemento adiacente. L'impostazione minima consentita è 1.01; non è prevista un'impostazione massima.

- **Consenti valori mancanti nelle condizioni.** `Vero` per consentire la verifica `IS MISSING` nelle regole.
- **Scarta risultati intermedi.** Se `Vero`, vengono restituiti solo i risultati finali della ricerca. Un risultato finale è un risultato che non viene ulteriormente rifinito nel processo di ricerca. Se `Falso`, vengono restituiti anche i risultati intermedi.

Personalizza campi disponibili: La finestra di dialogo Personalizza campi disponibili consente di selezionare i campi disponibili come attributi di segmento trovati dall'attività di mining.

Disponibile. Mostra i campi disponibili come attributi di segmenti. Per rimuovere campi dall'elenco, selezionare i campi desiderati e fare clic su **Rimuovi >>**. I campi selezionati vengono spostati dall'elenco Disponibile all'elenco Non disponibile.

Non disponibile. Mostra i campi che non sono disponibili come attributi di segmenti. Per includere i campi nell'elenco di quelli disponibili, selezionare i campi desiderati e fare clic su **<< Aggiungi**. I campi selezionati vengono spostati dall'elenco Non disponibile all'elenco Disponibile.

Organizzazione delle selezioni di dati: Mediante l'organizzazione delle selezioni di dati (di un insieme di dati di mining) è possibile specificare quali misure di valutazione devono essere analizzate da Visualizzatore dell'elenco di decisioni per trovare nuove regole e definire quali selezioni di dati devono essere utilizzate come base per le misure.

Per organizzare le selezioni di dati:

1. Dal menu Strumenti, scegliere **Organizza le selezioni di dati** oppure scegliere la stessa opzione facendo clic con il pulsante destro del mouse su un segmento. Viene visualizzata la finestra di dialogo Organizza le selezioni di dati.

   Nota: la finestra di dialogo Organizza selezioni dati consente anche di modificare o eliminare le selezioni di dati esistenti.
2. Fare clic sul pulsante **Aggiungi una nuova selezione di dati**. Nella tabella esistente viene aggiunta una nuova selezione di dati.
3. Fare clic su **Nome** e digitare un nome appropriato per la selezione.
4. Fare clic su **Partizione** e selezionare il tipo di partizione desiderato.
5. Fare clic su **Condizione** e selezionare un'opzione appropriata per la condizione. Quando si seleziona l'opzione **Specifica** viene aperta la finestra di dialogo Specifica condizione di selezione, contenente una serie di opzioni per la definizione di condizioni specifiche per i campi.
6. Definire la condizione appropriata e fare clic su **OK**.

Le selezioni di dati sono disponibili nell'elenco a discesa Selezione per creazione della finestra di dialogo Crea/Modifica attività di mining. L'elenco consente di selezionare la misura di valutazione utilizzata per una particolare attività di mining.

Regole di segmento

Le regole di segmento di modello si individuano eseguendo un'attività di mining basata su un modello di attività. È possibile aggiungere manualmente le regole di segmento ad un modello mediante la funzione Inserisci segmento o Modifica regola segmenti.

Se si decide di eseguire un'attività di mining per trovare nuove regole di segmento, gli eventuali risultati vengono visualizzati nella scheda Visualizzatore della finestra di dialogo Elenco interattivo. È possibile ridefinire rapidamente il modello selezionando uno dei risultati alternativi dalla finestra di dialogo Album modelli e facendo clic su **Carica**. In questo modo è possibile effettuare esperimenti con risultati diversi finché non si è pronti a creare un modello che descrive con precisione il gruppo obiettivo ideale.

Inserimento di segmenti: È possibile aggiungere manualmente le regole di segmento ad un modello mediante la funzione Inserisci segmento.

Per aggiungere una condizione di regola di segmento a un modello:

1. Nella finestra di dialogo Elenco interattivo, selezionare il punto del modello in cui si desidera aggiungere un nuovo segmento. Il nuovo segmento verrà inserito subito sopra il segmento selezionato.
2. Dal menu Modifica, scegliere **Inserisci segmento** oppure selezionare la stessa opzione facendo clic su un segmento con il pulsante destro del mouse.

   Viene visualizzata la finestra di dialogo Inserisci segmento, che consente di inserire nuove condizioni delle regole di segmento.
3. Fare clic su **Inserisci**. Viene visualizzata la finestra di dialogo Inserisci condizione, che consente di definire gli attributi per la nuova condizione delle regole.
4. Selezionare un campo e un operatore dagli elenchi a discesa.

   Nota: se viene selezionato l'operatore **Non in**, la condizione selezionata funzionerà come condizione di esclusione e viene visualizzata in rosso nella finestra di dialogo Inserisci regola. i Per esempio, la condizione `area geografica = 'CITTÀ'` visualizzata in rosso significa che `CITTÀ` è escluso dall'insieme di risultati.

5. Inserire uno o più valori o fare clic sull'icona **Inserisci valore** per visualizzare la finestra di dialogo Inserisci valore. La finestra di dialogo consente di selezionare un valore definito per il campo selezionato. Per esempio, il campo **sposato** presenterà i valori **sì** e **no**.
6. Scegliere **OK** per tornare alla finestra di dialogo Inserisci segmento. Fare clic su **OK** una seconda volta per aggiungere il segmento creato al modello.

Il nuovo segmento sarà visualizzato nel punto del modello specificato.

Modifica delle regole di segmento: La funzionalità Modifica regola di segmento consente di aggiungere, modificare o eliminare le condizioni delle regole dei segmenti.

Per modificare la condizione di una regola di segmento:

1. Selezionare il segmento di modello da modificare.
2. Dal menu Modifica scegliere **Modifica regola di segmento**, oppure selezionare la stessa opzione facendo clic sulla regola con il pulsante destro del mouse.

   Viene visualizzata la finestra di dialogo Modifica regola di segmento.
3. Selezionare la condizione appropriata e fare clic su **Modifica**.

   Viene visualizzata la finestra di dialogo Modifica condizione, che consente di definire gli attributi per la condizione della regola selezionata.
4. Selezionare un campo e un operatore dagli elenchi a discesa.

   Nota: se si seleziona l'operatore **Non in**, la condizione selezionata funzionerà come condizione di esclusione e viene visualizzata in rosso nella finestra di dialogo Modifica regola di segmento. Per esempio, la condizione `area geografica = 'CITTÀ'` visualizzata in rosso significa che `CITTÀ` è escluso dall'insieme di risultati.
5. Inserire uno o più valori o fare clic sul pulsante **Inserisci valore** per visualizzare la finestra di dialogo Inserisci valore. La finestra di dialogo consente di selezionare un valore definito per il campo selezionato. Per esempio, il campo **sposato** presenterà i valori **sì** e **no**.
6. Fare clic su **OK** per tornare alla finestra di dialogo Modifica regola di segmento. Fare clic su **OK** una seconda volta per tornare al modello di lavoro.

Il segmento selezionato sarà visualizzato con le condizioni della regola aggiornate.

Eliminazione delle condizioni delle regole di segmento: **Per eliminare la condizione di una regola di segmento:**

1. Selezionare il segmento di modello contenente le condizioni di regola da eliminare.
2. Dal menu Modifica scegliere **Modifica regola di segmento**, oppure selezionare la stessa opzione facendo clic sul segmento con il pulsante destro del mouse.

   Viene visualizzata la finestra di dialogo Modifica regola di segmento, che consente di eliminare una o più condizioni della regola di segmento.
3. Selezionare la condizione di regola appropriata e fare clic su **Elimina**.
4. Fare clic su **OK**.

L'eliminazione di una o più condizioni delle regole di segmento determina l'aggiornamento delle metriche delle misure nel riquadro Modello di lavoro.

Copia di segmenti: Visualizzatore dell'elenco di decisioni dispone di un comodo sistema per copiare i segmenti di modello. Se si desidera applicare un segmento di un modello a un altro modello è sufficiente copiare (o tagliare) il segmento da un modello e incollarlo in un altro modello. È possibile anche copiare un segmento da un modello visualizzato nel riquadro Anteprima alternative e incollarlo nel modello visualizzato nel riquadro Modello di lavoro. Queste funzioni di taglio, copia e incolla utilizzano gli Appunti di sistema per memorizzare o recuperare i dati temporanei. Pertanto, le condizioni e l'obiettivo vengono copiati negli Appunti. Il contenuto degli Appunti non è riservato all'uso in Visualizzatore

dell'elenco di decisioni ma può essere incollato anche in altre applicazioni. Per esempio, quando il contenuto degli Appunti viene incollato in un editor di testo, le condizioni e l'obiettivo vengono incollati in formato XML.

Per copiare o tagliare i segmenti di modello:

1. Selezionare il segmento di modello che si desidera utilizzare in un altro modello.
2. Dal menu Modifica scegliere **Copia** (o **Taglia)** oppure fare clic sul segmento di modello con il pulsante destro del mouse e selezionare **Copia** o **Taglia**.
3. Aprire il modello desiderato (in cui dovrà essere incollato il segmento di modello).
4. Selezionare uno dei segmenti di modello e fare clic su **Incolla**.

Nota: invece dei comandi **Taglia**, **Copia** e **Incolla**, è possibile utilizzare le combinazioni di tasti: **Ctrl+X**, **Ctrl+C** e **Ctrl+V**.

Il segmento copiato (o tagliato) viene inserito sopra il segmento di modello selezionato in precedenza. Le misure del segmento incollato e dei segmenti al di sotto di esso vengono ricalcolate.

Nota: entrambi i modelli in questa procedura devono essere basati sullo stesso modello sottostante e contenere lo stesso obiettivo; in caso contrario, viene visualizzato un messaggio di errore.

Modelli alternativi: Se ci sono più risultati, la scheda Alternative visualizza i risultati di ogni attività di mining. Ogni risultato è costituito dalle condizioni dei dati selezionati che risultano più simili all'obiettivo, nonché da tutte le alternative sufficientemente soddisfacenti. Il numero totale di alternative visualizzate dipende dai criteri di ricerca usati durante il processo di analisi.

Per visualizzare i modelli alternativi:

1. Fare clic su un modello alternativo nella scheda Alternative. I segmenti di modello alternativo vengono visualizzati o sostituiscono quelli di modello corrente nel riquadro Anteprima alternative.
2. Per utilizzare un modello alternativo nel riquadro Modello di lavoro, selezionare il modello e fare clic su **Carica** nel riquadro Anteprima alternative o fare clic con il pulsante destro del mouse sul nome di un'alternativa nella scheda Alternative e selezionare **Carica**.

Nota: i modelli alternativi non vengono salvati quando si genera un nuovo modello.

Personalizzazione di un modello

I dati non sono statici: i clienti si spostano, si sposano e cambiano lavoro. I prodotti perdono il focus di mercato e diventano obsoleti.

Visualizzatore dell'elenco di decisioni offre agli utenti di business la flessibilità necessaria per adattare facilmente e rapidamente i modelli alle nuove situazioni. È possibile cambiare un modello modificando, stabilendo la priorità, eliminando o disattivando segmenti di modello specifici.

Definizione delle priorità tra i segmenti: È possibile classificare le regole dei modelli nell'ordine desiderato. Per default, i segmenti di modello vengono visualizzati in ordine di priorità (il segmento con la massima priorità è visualizzato per primo). Quando si assegna una priorità diversa a uno o più segmenti, il modello viene modificato di conseguenza. È possibile modificare il modello in funzione delle proprie esigenze spostando i segmenti in una posizione di maggiore o minore priorità.

Per definire le priorità dei segmenti di modello:

1. Selezionare il segmento di modello a cui si desidera assegnare una priorità diversa.
2. Fare clic su uno dei due pulsanti con le frecce nella barra degli strumenti del riquadro Modello di lavoro per spostare il segmento di modello selezionato verso l'alto o verso il basso nell'elenco.

Dopo la definizione delle priorità, tutti i risultati delle precedenti valutazioni vengono ricalcolati e vengono visualizzati i nuovi valori.

Eliminazione di segmenti: Per eliminare uno o più segmenti:

1. Selezionare un segmento di modello.
2. Scegliere **Elimina segmento** dal menu Modifica o fare clic sul pulsante di eliminazione nella barra degli strumenti del riquadro Modello di lavoro.

Le misure del modello modificato vengono ricalcolate e il modello viene modificato di conseguenza.

Esclusione di segmenti: Poiché la ricerca riguarda gruppi particolari, le azioni di business saranno basate probabilmente su una selezione dei segmenti di modello. Quando si esegue la distribuzione di un modello è possibile decidere di escludere dei segmenti dal modello. Il calcolo del punteggio assegna ai segmenti esclusi dei valori null. L'esclusione di un segmento non ne comporta il mancato utilizzo; semplicemente, tutti i record corrispondenti alla regola specificata vengono esclusi dalla mailing list. La regola continua a essere applicata, ma in maniera diversa.

Per escludere segmenti di modello specifici:

1. Selezionare un segmento dal riquadro Modello di lavoro.
2. Fare clic sul pulsante **Attiva/Disattiva esclusione segmento** nella barra degli strumenti del riquadro Modello di lavoro. Nella colonna Obiettivo selezionata del segmento è ora visualizzata la dicitura **Esclusa**.

Nota: al contrario dei segmenti eliminati, i segmenti esclusi restano disponibili per essere riutilizzati nel modello finale. I segmenti esclusi influiscono sui risultati dei grafici.

Cambia valore obiettivo: La finestra di dialogo Cambia valore obiettivo consente di modificare il valore obiettivo per il campo obiettivo corrente.

Le snapshot e i risultati della sessione con un valore obiettivo diverso da quello del modello di lavoro si distinguono per il colore di sfondo della tabella, che per quella riga diventa giallo. Questo indica che la snapshot/il risultato della sessione è superato.

Nella finestra di dialogo **Crea/Modifica attività di mining** è visualizzato il valore obiettivo per il modello di lavoro corrente. Il valore obiettivo non viene salvato insieme all'attività di mining, ma viene invece ricavato dal valore del modello di lavoro.

Quando si innalza a modello di lavoro un modello salvato con un valore obiettivo diverso da quello del modello di lavoro corrente (per esempio modificando un risultato alternativo o una copia di una snapshot), il valore obiettivo del modello salvato viene impostato sullo stesso valore di quello del modello di lavoro (il valore obiettivo visualizzato nel riquadro Modello di lavoro resta invariato). Le metriche del modello vengono rivalutate con il nuovo obiettivo.

Genera nuovo modello

La finestra di dialogo Genera nuovo modello contiene una serie di opzioni per assegnare un nome al modello e selezionare la posizione in cui deve essere creato il nuovo nodo.

Nome modello. Selezionare **Personalizzato** per modificare il nome generato automaticamente o per creare un nome univoco per il nodo, come visualizzato nell'area del flusso.

Crea nodo in. Selezionare **Area** per collocare il nuovo modello nell'area di lavoro; selezionare **Palette MG** per collocare il nuovo modello nella palette Modelli; selezionare **Entrambe** per collocare il nuovo modello sia nell'area di lavoro, sia nella palette Modelli.

Includi stato di sessione interattiva. Quando questa opzione è attivata, lo stato della sessione interattiva viene mantenuto nel modello generato. Quando in seguito si genera un nodo Modelli a partire dal modello, lo stato viene mantenuto e utilizzato per inizializzare la sessione interattiva. Indipendentemente dal fatto che l'opzione sia selezionata o meno, il modello stesso calcola i punteggi dei nuovi dati in modo identico. Quando l'opzione non è selezionata, il modello è comunque in grado di generare un nodo di creazione, ma si tratterà di un nodo più generico che avvia una nuova sessione interattiva anziché riprendere dal punto in cui si era interrotta la sessione precedente. Se si modificano le impostazioni del nodo ma lo si esegue con uno stato salvato, le impostazioni modificate vengono ignorate a favore di quelle dello stato salvato.

Nota: le metriche standard sono le uniche metriche che rimangono nel modello. Le metriche aggiuntive vengono mantenute insieme allo stato interattivo. Il modello generato non rappresenta lo stato interattivo salvato dell'attività di mining. Quando si avvia Visualizzatore dell'elenco di decisioni, il programma visualizza le impostazioni definite originariamente tramite il Visualizzatore.

Per ulteriori informazioni, consultare l'argomento “Rigenerazione di un nodo Modelli” a pagina 49.

Valutazione del modello

Perché la modellazione sia corretta è necessaria un'attenta valutazione del modello prima della sua effettiva implementazione nell'ambiente di produzione. Visualizzatore dell'elenco di decisioni offre numerose misure statistiche e di business utilizzabili per valutare l'impatto di un modello nel mondo reale. Tra queste, grafici dei profitti e interoperabilità completa con Excel, che consentono la simulazione di scenari per l'analisi dei costi/benefici al fine di valutare l'impatto della distribuzione.

Il modello può essere valutato nei modi seguenti:

- Utilizzando le misure del modello statistiche e di business predefinite disponibili in Visualizzatore dell'elenco di decisioni (probabilità, frequenza).
- Valutando misure importate da Microsoft Excel.
- Visualizzando il modello mediante un grafico dei profitti.

Organizzazione delle misure del modello: Visualizzatore dell'elenco di decisioni è dotato di una serie di opzioni per la definizione delle misure che vengono calcolate e visualizzate sotto forma di colonne. Ogni segmento può comprendere la copertura, la frequenza, la probabilità di default e le misure errate rappresentate come colonne. È anche possibile creare nuove misure che saranno visualizzate sotto forma di colonne.

Definizione delle misure del modello

Per aggiungere una misura al modello o per definirne una esistente:

1. Scegliere **Organizza misure del modello** dal menu Strumenti o fare clic sul modello con il pulsante destro del mouse per selezionare la stessa opzione. Viene visualizzata la finestra di dialogo Organizza misure del modello.
2. Fare clic sul pulsante **Aggiungi una nuova misura del modello** (a destra della colonna Mostra). Nella tabella viene visualizzata una nuova misura.
3. Assegnare un nome alla misura e scegliere il tipo, l'opzione di visualizzazione e la selezione desiderati. La colonna Mostra indica se la misura sarà visualizzata per il modello di lavoro. Quando si definisce una misura esistente, scegliere una metrica e una selezione appropriate e indicare se la misura dovrà essere visualizzata per il modello di lavoro.
4. Fare clic su **OK** per tornare allo spazio di lavoro di Visualizzatore dell'elenco di decisioni. Se la colonna Mostra della nuova misura è stata selezionata, la nuova misura sarà visualizzata per il modello di lavoro.

Metriche personalizzate in Excel

Per ulteriori informazioni, consultare l'argomento “Valutazione in Excel”.

Aggiornamento delle misure: In alcuni casi può essere necessario ricalcolare le misure del modello, per esempio quando si applica un modello esistente a un nuovo insieme di clienti.

Per ricalcolare (aggiornare) le misure del modello:

Scegliere **Aggiorna tutte le misure** dal menu Modifica.

oppure

Premere F5.

Tutte le misure vengono ricalcolate e per il modello di lavoro vengono visualizzati i nuovi valori.

Valutazione in Excel: È possibile integrare Visualizzatore dell'elenco di decisioni con Microsoft Excel per utilizzare i propri calcoli dei valori e le formule dei profitti direttamente all'interno del processo di creazione del modello in modo da simulare scenari di costi/benefici. Il collegamento con Excel consente di esportare i dati in Excel, dove è possibile utilizzarli per creare grafici di presentazione, calcolare misure personalizzate, come misure ROI e di profitto complesse, e visualizzarli in Visualizzatore dell'elenco di decisioni durante la creazione del modello.

Nota: per consentire l'utilizzo di un foglio di calcolo di Excel, l'esperto di analisi CRM dovrà definire i dati di configurazione per sincronizzare Visualizzatore dell'elenco di decisioni con Microsoft Excel. La configurazione, contenuta in un file di foglio di calcolo di Excel, indica quali dati vengono trasferiti da Visualizzatore dell'elenco di decisioni a Excel e viceversa.

La procedura descritta di seguito è valida solo se è installato MS Excel. Se Excel non è installato, le opzioni per la sincronizzazione dei modelli con Excel non sono visualizzate.

Per sincronizzare i modelli con MS Excel:

1. Aprire il modello, eseguire una sessione interattiva e scegliere **Organizza misure del modello** dal menu Strumenti.
2. Selezionare **Sì** per l'opzione **Calcola le misure personalizzate in Excel**. Viene attivato il campo **Cartella di lavoro**, che consente di selezionare un modello di cartella di lavoro Excel preconfigurato.
3. Fare clic sul pulsante **Collega a Excel**. Viene visualizzata la finestra di dialogo Apri, che consente di passare alla directory del modello preconfigurato del file system locale o di rete.
4. Selezionare il modello di Excel appropriato e fare clic su **Apri**. Il modello di Excel selezionato viene avviato; utilizzare la barra delle applicazioni di Windows (oppure premere Alt-Tab) per ritornare alla finestra di dialogo Scegli input per misure personalizzate.
5. Selezionare le mappature appropriate fra i nomi delle metriche definiti nel modello di Excel e quelli delle metriche del modello e fare clic su **OK**.

Una volta stabilito il collegamento, Excel viene avviato con il modello predefinito che visualizza un foglio di calcolo con le regole del modello. I risultati calcolati in Excel vengono visualizzati come nuove colonne in Visualizzatore dell'elenco di decisioni.

Nota: le metriche di Excel non vengono conservate al salvataggio del modello; le metriche sono valide solo durante la sessione attiva. È tuttavia possibile creare delle snapshot che includano le metriche di Excel. Tali metriche salvate nelle snapshot sono valide solo ai fini del confronto cronologico e non vengono aggiornate alla riapertura. Per ulteriori informazioni, consultare l'argomento “Scheda Snapshot” a pagina 147. Le metriche di Excel non verranno visualizzate nelle snapshot fino a quando non viene nuovamente stabilita una connessione al modello Excel.

Impostazione dell'integrazione con MS Excel: L'integrazione fra Visualizzatore dell'elenco di decisioni e Microsoft Excel viene attuata con l'uso di un modello di foglio di calcolo di Excel predefinito. Il modello è composto da tre fogli di lavoro:

Misure del modello. Visualizza le misure importate da Visualizzatore dell'elenco di decisioni, le misure personalizzate di Excel e i totali dei calcoli (definiti nel foglio di lavoro Impostazioni).

Impostazioni. Fornisce le variabili per generare i calcoli in base alle misure importate da Visualizzatore dell'elenco di decisioni e alle misure personalizzate di Excel.

Configurazione. Fornisce le opzioni necessarie per indicare quali misure devono essere importate da Visualizzatore dell'elenco di decisioni e per definire le misure personalizzate di Excel.

AVVISO: la struttura del foglio di lavoro Configurazione è definita in modo rigido. **NON** modificare le celle nell'area ombreggiata verde.

- **Metriche dal modello.** Indica quali metriche di Visualizzatore dell'elenco di decisioni vengono utilizzate nei calcoli.
- **Metriche al modello.** Indica quali metriche generate da Excel saranno restituite a Visualizzatore dell'elenco di decisioni. Le metriche generate da Excel sono visualizzate come colonne di nuove misure in Visualizzatore dell'elenco di decisioni.

Nota: le metriche di Excel sono valide solo durante la sessione attiva e non vengono conservate insieme al modello quando si genera un nuovo modello.

Modifica delle misure del modello: Gli esempi riportati di seguito spiegano come modificare le misure del modello in vari modi:

- Modifica di una misura esistente.
- Importazione di una misura standard aggiuntiva dal modello.
- Esportazione di una misura personalizzata aggiuntiva nel modello.

Modifica di una misura esistente

1. Aprire il modello e selezionare il foglio di lavoro Configurazione.
2. Modificare il **Nome** o la **Descrizione** evidenziandoli e sovrascrivendoli.

   Notare che, se si desidera modificare una misura, ad esempio per richiedere all'utente la Probabilità invece della Frequenza, è necessario solo modificare il nome e la descrizione in **Metriche dal modello** – la modifica viene quindi visualizzata nel modello e l'utente può scegliere la misura appropriata da mappare.

Importazione di una misura standard aggiuntiva dal modello

1. Aprire il modello e selezionare il foglio di lavoro Configurazione.
2. Dai menu, scegliere:

   Strumenti > **Protezione** > **Rimuovi protezione foglio**
3. Selezionare la cella A5, che è ombreggiata gialla e contiene la parola **Fine**.
4. Dai menu, scegliere:

   Inserisci > **Righe**
5. Immettere il **Nome** e la **Descrizione** della nuova misura. Per esempio, **Errore** diventa **Errore associato al segmento**.
6. Nella cella C5, immettere la formula **=COLUMN('Model Measures'!N3)**.
7. Nella cella D5, immettere la formula **=ROW('Model Measures'!N3)+1**.

   Queste formule determinano la visualizzazione della nuova misura nella colonna N del foglio di lavoro Misure del modello, che attualmente è vuoto.

8. Dai menu, scegliere:

   Strumenti > **Protezione** > **Proteggi foglio**
9. Fare clic su **OK**.
10. Nel foglio di lavoro Misure del modello, assicurarsi che nella cella N3 il titolo della nuova colonna sia **Errore**.
11. Selezionare l'intera colonna N.
12. Dai menu, scegliere:

    Formato > **Celle**
13. Per default, tutte le celle hanno una categoria numerica **Generale**. Fare clic su **Percentuale** per modificare la modalità di visualizzazione delle cifre. Ciò consente di verificare i dati in Excel; inoltre, permette di utilizzare i dati in altri modi, per esempio come output di un grafico.
14. Fare clic su **OK**.
15. Salvare il foglio di calcolo come modello Excel 2003, con un nome univoco e l'estensione file *.xlt*. Per facilitare l'individuazione del nuovo modello, si consiglia di salvarlo nella directory predefinita dei modelli nel file system locale o di rete.

Esportazione di una misura personalizzata aggiuntiva nel modello

1. Aprire il modello al quale è stata aggiunta la colonna Errore nell'esempio precedente e selezionare il foglio di lavoro Configurazione.
2. Dai menu, scegliere:

   Strumenti > **Protezione** > **Rimuovi protezione foglio**
3. Selezionare la cella A14, che è ombreggiata gialla e contiene la parola **Fine**.
4. Dai menu, scegliere:

   Inserisci > **Righe**
5. Immettere il **Nome** e la **Descrizione** della nuova misura. Per esempio, **Errore scalato** e **Scala applicata a errore da Excel**.
6. Nella cella C14, immettere la formula **=COLUMN('Model Measures'!O3)**.
7. Nella cella D14, immettere la formula **=ROW('Model Measures'!O3)+1**.

   Queste formule specificano che la colonna O fornirà la nuova misura al modello.
8. Selezionare il foglio di lavoro Impostazioni.
9. Nella cella A17, immettere la descrizione **'- Errore scalato**.
10. Nella cella B17, immettere il fattore di scala **10**.
11. Nel foglio di lavoro Misure del modello, immettere la descrizione **Errore scalato** nella cella O3 come titolo della nuova colonna.
12. Nella cella O4, immettere la formula **=N4*Settings!B17**.
13. Selezionare l'angolo della cella O4 e trascinarlo fino alla cella O22 per copiare la formula in ogni cella.
14. Dai menu, scegliere:

    Strumenti > **Protezione** > **Proteggi foglio**
15. Fare clic su **OK**.
16. Salvare il foglio di calcolo come modello Excel 2003, con un nome univoco e l'estensione file *.xlt*. Per facilitare l'individuazione del nuovo modello, si consiglia di salvarlo nella directory predefinita dei modelli nel file system locale o di rete.

Quando ci si connette a Excel utilizzando questo modello, il valore Errore è disponibile come nuova misura personalizzata.

Visualizzazione dei modelli

Il modo migliore per comprendere l'impatto di un modello è visualizzarlo. Con l'uso di un grafico dei profitti è possibile ottenere una dettagliata visione dei vantaggi di business e tecnici del proprio modello studiando l'effetto di più alternative in tempo reale. La sezione “Grafico dei profitti” mostra i vantaggi dell'utilizzo di un modello rispetto ad un processo decisionale casuale e consente il confronto diretto di più grafici quando sono disponibili modelli alternativi.

Grafico dei profitti: Nel grafico dei profitti sono rappresentati i valori della colonna *Guadagno %* della tabella. I guadagni sono definiti come la proporzione di risultati in ogni incremento rispetto al numero totale di risultati nella struttura ad albero, mediante l'equazione seguente:

(risultati in incremento /numero totale di risultati) x 100%

I grafici dei profitti illustrano la modalità di progettazione più idonea per identificare una percentuale specifica di tutti i risultati nella struttura ad albero. La linea diagonale rappresenta la risposta prevista per l'intero campione, se il modello non viene utilizzato. In questo caso, la percentuale di risposta è costante, poiché la probabilità che un utente risponda è uguale a quella di un altro utente. Per raddoppiare le possibilità, sarebbe necessario formulare la domanda a un numero doppio di persone. La linea curva indica di quanto è possibile aumentare la risposta inserendo solo le persone classificate nei percentili più alti in base al guadagno. Per esempio, se si include il 50% superiore, si può ottenere un incremento netto del 70 percento di risposte positive. Maggiore è la curvatura, maggiore è il guadagno.

Per visualizzare un grafico dei profitti:

1. Aprire un flusso contenente un nodo Elenco di decisioni e avviare una sessione interattiva dal nodo.
2. Fare clic sulla scheda **Guadagni**. A seconda delle partizioni specificate, i grafici visualizzati possono essere uno o due (due grafici vengono visualizzati, per esempio, quando per le misure del modello vengono definite sia la partizione di addestramento che quella di test).

Per default, i grafici sono visualizzati come segmenti. Per passare alla visualizzazione dei grafici come quantili, selezionare **Quantili** e quindi selezionare il metodo quantile appropriato dal menu a discesa.

Grafici: opzioni: La funzione Opzioni grafico fornisce una serie di opzioni per selezionare i modelli e le snapshot da rappresentare graficamente e le partizioni da includere nel grafico, nonché per decidere se visualizzare o meno le etichette dei segmenti.

Modelli per il plot

Modelli correnti. Consente di selezionare i modelli da rappresentare graficamente. È possibile selezionare il modello di lavoro o qualsiasi modello di snapshot creato.

Partizioni per il plot

Partizioni per il grafico a sinistra. L'elenco a discesa contiene una serie di opzioni per visualizzare tutte le partizioni definite o tutti i dati.

Partizioni per il grafico a destra. L'elenco a discesa contiene una serie di opzioni per visualizzare tutte le partizioni definite, tutti i dati o solo il grafico a sinistra. Quando si seleziona **Grafico solo a sinistra** viene visualizzato solo il grafico a sinistra.

Visualizza etichette dei segmenti. Quando questa opzione è selezionata, le etichette dei singoli segmenti vengono visualizzate nei grafici.

Capitolo 10. Modelli statistici

Nei modelli statistici vengono utilizzate equazioni matematiche per codificare le informazioni estratte dai dati. In alcuni casi le tecniche di modellazione statistica possono fornire modelli adeguati molto velocemente. Anche per problemi per cui si possono ottenere risultati migliori con tecniche di apprendimento automatico più flessibili (quali le reti neurali), è possibile utilizzare alcuni modelli statistici come modelli predittivi di riferimento per valutare le prestazioni di tecniche più avanzate.

Sono disponibili i seguenti nodi Modelli statistici.

I modelli di regressione lineare prevedono un target continuo basato sulle relazioni lineari tra l'obiettivo e uno o più predittori.

La regressione logistica, una tecnica statistica che consente di classificare i record in base ai valori dei campi di input, è analoga alla regressione lineare ma, al posto di un intervallo numerico, prende un campo obiettivo categoriale.

Il nodo fattoriale/PCA offre potenti tecniche di riduzione dei dati che consentono di diminuirne la complessità. L'analisi dei componenti principali (PCA, Principal Components Analysis) trova le combinazioni lineari dei campi di input che catturano meglio la varianza nell'intero insieme di campi, dove i componenti sono ortogonali (perpendicolari) l'uno rispetto all'altro. L'analisi fattoriale tenta di identificare i concetti sottostanti, o fattori, che spiegano lo schema delle correlazioni all'interno dell'insieme di campi osservati. Entrambi gli approcci mirano a trovare un numero ridotto di campi derivati che riassumono in modo efficace le informazioni presenti nell'insieme originale di campi.

L'analisi discriminante prevede presupposti più rigidi rispetto alla regressione logistica, ma può essere una valida alternativa o un complemento dell'analisi di regressione logistica quando vengono soddisfatti tali presupposti.

Il modello Lineare generalizzato amplia il modello lineare generale in modo che la variabile dipendente venga linearmente correlata ai fattori e alle covariate tramite una funzione di collegamento specifica. Inoltre, il modello consente alla variabile dipendente di avere una distribuzione non normale. Copre la funzionalità di un grande numero di modelli statistici, inclusi modelli di regressione lineare, modelli di regressione logistica, modelli loglineari per dati dei conteggi e modelli di sopravvivenza censurati per intervallo.

Un modello misto lineare generalizzato (GLMM) estende il modello lineare in modo che l'obiettivo possa avere una distribuzione non normale, sia linearmente correlato ai fattori e alle covariate tramite una funzione di collegamento specifica e in modo che le osservazioni possano essere correlate. I modelli misti lineari generalizzati includono un'ampia gamma di modelli, dalla regressione lineare semplice ai modelli multilivello complessi per i dati longitudinali non normali.

Il nodo Regressione di Cox consente di generare un modello di sopravvivenza per i dati della relazione tempo-evento in presenza di record censurati. Il modello produce una funzione di sopravvivenza che prevede la probabilità che l'evento di interesse si sia verificato a una determinata ora (*t*) per i valori dati delle variabili di input.

Nodo lineare

La regressione lineare è una tecnica statistica comune che consente di classificare i record in base ai valori dei campi di input numerici. La regressione lineare rappresenta una linea retta o un piano che riduce al minimo le differenze tra i valori di output previsti e quelli effettivi.

Requisiti. In un modello di regressione lineare è possibile utilizzare solo campi numerici. È necessario disporre di un solo campo obiettivo (con il ruolo impostato su *Obiettivo*) e uno o più predittori (con il ruolo impostato su *Input*). I campi con ruolo *Entrambi* o *Nessuno* vengono ignorati, così come vengono ignorati i campi non numerici. Se necessario, è possibile ricodificare i campi non numerici utilizzando un nodo Ricava.

Efficacia. I modelli di regressione lineare sono relativamente semplici e danno una formula matematica di facile interpretazione per la generazione delle previsioni. La regressione lineare è una procedura statistica nota da tempo, pertanto le proprietà di questi modelli non presentano difficoltà di comprensione. Inoltre, in genere l'addestramento dei modelli lineari è rapido. Il nodo lineare fornisce metodi di selezione automatica dei campi per eliminare dall'equazione i campi di input non significativi.

Nota: nei casi in cui il campo obiettivo sia categoriale anziché essere un intervallo continuo, per esempio *sì/no* o *churn/don't churn*, come alternativa è possibile utilizzare la regressione logistica. La regressione logistica inoltre supporta gli input non numerici, eliminando la necessità di ricodificare questi campi. Per ulteriori informazioni, consultare l'argomento "Nodo Logistica" a pagina 169.

Modelli lineari

I modelli lineari prevedono un target continuo basato sulle relazioni lineari tra l'obiettivo e uno o più predittori.

I modelli lineari sono relativamente semplici e forniscono una formula matematica di facile interpretazione per il calcolo del punteggio. Le proprietà di questi modelli sono di facile comprensione e normalmente vengono creati più rapidamente rispetto ad altri tipi di modelli (quali le reti neurali o le strutture ad albero delle decisioni) dello stesso insieme di dati.

Esempio. Al fine di indagare sulle richieste di indennizzo dei proprietari di immobili, una compagnia di assicurazioni con risorse limitate desidera creare un modello per stimare i costi delle richieste di risarcimento. Distribuendo tale modello ai centri servizi, i rappresentanti possono immettere le informazioni sulle richieste mentre sono al telefono con un cliente ed ottenere immediatamente il costo "previsto" della richiesta di risarcimento in base ai dati precedenti. Per ulteriori informazioni, consultare l'argomento .

Requisiti dei campi. Devono essere presenti un obiettivo ed almeno un input. Per default, i campi con ruoli predefiniti Entrambi o Nessuno non vengono utilizzati. L'obiettivo deve essere continuo (scala). Non sono previste limitazioni per il livello di misurazione sui predittori (input); i campi relativi alla categoria (indicatori, nominali ed ordinali) sono utilizzati come fattori nel modello ed i campi continui sono utilizzati come covariate.

Obiettivi

Come si desidera procedere?

- **Creare un nuovo modello.** Creare un modello completamente nuovo. Questa è la procedura normale per il nodo.
- **Continuare l'addestramento di un modello già esistente.** L'addestramento continua con l'ultimo modello generato correttamente dal nodo. In questo modo sarà possibile aggiornare un modello esistente senza dover accedere ai dati originali, aumentando così significativamente le prestazioni

poiché nel flusso verranno utilizzati solo i record nuovi o aggiornati. I dettagli del modello precedente vengono archiviati con il nodo Modelli, consentendo di utilizzare questa opzione anche se il nugget del modello precedente non è più disponibile nel flusso o nella palette Modelli.

*Nota:*quando questa opzione è abilitata, tutti gli altri controlli nelle schede Opzioni di creazione e dei campi sono disabilitati.

Qual è l'obiettivo principale? Selezionare l'obiettivo appropriato.

- **Creare un modello standard.** Il metodo crea un unico modello per la previsione dell'obiettivo utilizzando i predittori. In generale, i modelli standard sono più semplici da interpretare e consentono un più rapido calcolo del punteggio rispetto agli insiemi utilizzati per boosting, bagging o insiemi di dati di grandi dimensioni.
- **Migliorare l'accuratezza del modello (boosting).** Il metodo crea un modello di classificazione binario mediante la tecnica di boosting, che genera una sequenza di modelli per ottenere previsioni più precise. La creazione e il calcolo del punteggio degli insiemi può richiedere più tempo rispetto al modello standard.

  Il boosting genera una successione di "modelli di componenti", ciascuno dei quali viene creato a partire dall'intero insieme di dati. Prima di creare ogni successivo modello di componente, i record vengono ponderati in base ai residui del modello di componente precedente. Ai casi con molti residui viene attribuito un peso di analisi relativamente superiore per far sì che il modello di componente successivo si concentri sulla corretta previsione di quei record. Insieme, questi modelli di componenti formano un modello di classificazione binario. Tale modello calcola il punteggio dei nuovi record utilizzando una regola di combinazione; le regole disponibili dipendono dal livello di misurazione dell'obiettivo.
- **Migliorare la stabilità del modello (bagging).** Il metodo crea un modello di classificazione binario mediante la tecnica di bagging (bootstrap aggregating), che genera più modelli per ottenere previsioni più affidabili. La creazione e il calcolo del punteggio degli insiemi può richiedere più tempo rispetto al modello standard.

  L'aggregazione bootstrap (bagging) genera repliche dell'insieme di dati di addestramento mediante il campionamento con sostituzione dall'insieme di dati originale. Questa operazione crea campioni di bootstrap di dimensioni uguali a quelle dell'insieme di dati originale. Su ogni replica viene quindi creato un "modello di componente". Insieme, questi modelli di componenti formano un modello di classificazione binario. Tale modello calcola il punteggio dei nuovi record utilizzando una regola di combinazione; le regole disponibili dipendono dal livello di misurazione dell'obiettivo.
- **Creare un modello per dataset di grandi dimensioni (richiede IBM SPSS Modeler Server).** Il metodo crea un modello di classificazione binario suddividendo l'insieme di dati in blocchi di dati separati. Scegliere questa opzione se l'insieme di dati è troppo grande per creare uno dei modelli descritti sopra o per la creazione incrementale del modello. Questa opzione può richiedere meno tempo per la creazione, ma più tempo per il calcolo del punteggio rispetto al modello standard. Questa opzione richiede la connettività IBM SPSS Modeler Server .

Consultare “Insiemi” a pagina 165 per le impostazioni relative a boosting, bagging ed insiemi di dati di grandi dimensioni.

Di base

Prepara automaticamente i dati. Questa opzione consente alla procedura di trasformare internamente l'obiettivo e i predittori in modo tale da ottimizzare il potere predittivo del modello; eventuali trasformazioni vengono salvate con il modello e applicate ai nuovi dati per il calcolo del punteggio. Le versioni originali dei campi trasformati vengono escluse dal modello. Per default, vengono eseguite le seguenti operazioni di preparazione automatica dei dati.

- **Gestione di data e ora.** Ciascun predittore di date viene trasformato in un nuovo predittore continuo che contiene il tempo trascorso a partire da una data di riferimento (1970-01-01). Ogni predittore di ora viene trasformato in un nuovo predittore continuo contenente il tempo trascorso a partire da un orario di riferimento (00:00:00).

- **Regola livello di misurazione.** I predittori continui con meno di 5 valori distinti vengono riformulati come predittori ordinali. I predittori ordinali con più di 10 valori distinti vengono riformulati come predittori continui.
- **Gestione dei valori anomali.** I valori dei predittori continui che superano un valore di interruzione (3 deviazioni standard dalla media) vengono impostati sul valore di interruzione.
- **Gestione dei valori mancanti.** I valori mancanti dei predittori nominali vengono sostituiti con la modalità della partizione di addestramento. I valori mancanti dei predittori ordinali vengono sostituiti con la mediana della partizione di addestramento. I valori mancanti dei predittori continui vengono sostituiti con la media della partizione di addestramento.
- **Unione supervisionata.** Questa opzione consente di creare un modello più facilmente gestibile riducendo il numero dei campi da elaborare in associazione con l'obiettivo. Le categorie simili vengono identificate in base alla relazione tra input e obiettivo. Le categorie che non presentano differenze significative (ovvero che hanno un valore P superiore a 0,1) vengono unite. Se tutte le categorie vengono unite in una, la versione originale e quella derivata del campo vengono escluse dal modello perché non hanno un valore come predittore.

Livello di confidenza. Si tratta del livello di confidenza utilizzato per calcolare stime di intervallo per i coefficienti del modello nella visualizzazione Coefficienti. Specificare un valore maggiore di 0 e minore di 100. Il valore di default è 95.

Selezione modello

Metodo di selezione del modello. Scegliere uno dei metodi di selezione del modello descritti di seguito o **Includi tutti i predittori**, che immette tutti i predittori disponibili come termini di modello di effetti principali. Per default si utilizza il metodo **Stepwise in avanti**.

Selezione Stepwise in avanti. La selezione viene avviata senza effetti nel modello e aggiunge o elimina effetti un passaggio alla volta finché non vi sono più effetti da aggiungere o eliminare, in base ai criteri stepwise.

- **Criteri per immissione/eliminazione.** Statistica utilizzata per determinare se un effetto deve essere aggiunto o eliminato dal modello. **Il criterio di informazione (AICC)** si basa sulla probabilità del set di addestramento in relazione al modello e viene adeguato in modo da penalizzare i modelli eccessivamente complessi. **Statistiche F** si basa su un test statistico del miglioramento nell'errore del modello. **R-quadrato corretto** si basa sull'adattamento del set di addestramento e viene adeguato in modo da penalizzare i modelli eccessivamente complessi. **Il criterio di prevenzione del sovradattamento (ASE)** si basa sull'adattamento (errore quadratico medio o ASE) dell'insieme di prevenzione del sovradattamento. L'insieme di prevenzione del sovradattamento è un sottocampione casuale di circa il 30% dell'insieme di dati originale non utilizzato per l'addestramento del modello.

  Se si sceglie un criterio diverso da **Statistiche F**, in ogni passaggio l'effetto che corrisponde al massimo aumento positivo nel criterio viene aggiunto al modello. Tutti gli effetti presenti nel modello che corrispondono a una diminuzione nel criterio vengono eliminati.

  Se si sceglie il criterio **Statistiche F**, in ogni passaggio l'effetto che ha il valore *p* più piccolo inferiore alla soglia specificata, **Includi effetti con valori P minori di**, viene aggiunto al modello. Il valore di default è 0.05. Tutti gli effetti presenti nel modello che hanno un valore *p* superiore alla soglia specificata, **Elimina effetti con valori P maggiori di**, vengono eliminati. Il valore predefinito è 0.10.
- **Personalizza il numero massimo di effetti nel modello finale.** Per default, tutti gli effetti disponibili possono essere immessi nel modello. In alternativa, se l'algoritmo stepwise conclude un passaggio con il numero massimo di effetti specificato, l'algoritmo si interrompe in corrispondenza dell'insieme di effetti corrente.
- **Personalizza il numero massimo di fasi.** L'algoritmo stepwise si interrompe dopo un determinato numero di passaggi. Per default, il numero è 3 volte il numero di effetti disponibili. In alternativa, specificare un numero intero positivo come numero massimo di passaggi.

Selezione dei sottoinsiemi migliori. Verifica tutti i modelli "possibili" o almeno un sottoinsieme di modelli possibili più grande rispetto al metodo stepwise in avanti, in modo tale da scegliere il migliore in

assoluto in base al criterio di selezione dei sottoinsiemi migliori. **Il criterio di informazione (AICC)** si basa sulla probabilità del set di addestramento in relazione al modello e viene adeguato in modo da penalizzare i modelli eccessivamente complessi. **R-quadrato corretto** si basa sull'adattamento del set di addestramento e viene adeguato in modo da penalizzare i modelli eccessivamente complessi. **Il criterio di prevenzione del sovradattamento (ASE)** si basa sull'adattamento (errore quadratico medio o ASE) dell'insieme di prevenzione del sovradattamento. L'insieme di prevenzione del sovradattamento è un sottocampione casuale di circa il 30% dell'insieme di dati originale non utilizzato per l'addestramento del modello.

Il modello con il massimo valore del criterio viene scelto come modello migliore.

Nota: la selezione dei sottoinsiemi migliori richiede un maggior numero di operazioni di calcolo rispetto alla selezione stepwise in avanti. Quando la sezione dei sottoinsiemi migliori viene eseguita in associazione a boosting, bagging o insiemi di dati di grandi dimensioni, la creazione può richiedere molto più tempo rispetto alla creazione di un modello standard con la selezione stepwise in avanti.

Insiemi

Queste impostazioni determinano il comportamento dei classificatori binari che si verificano quando negli obiettivi sono richiesti boosting, bagging o insiemi di dati di grandi dimensioni. Le opzioni non applicate all'obiettivo selezionato vengono ignorate.

Bagging e insiemi di dati di grandi dimensioni. Quando si calcola il punteggio di un insieme, questo tipo di regola consente di combinare i valori previsti provenienti dai modelli di base per calcolare il valore del punteggio dell'insieme.

- **Regola di combinazione di default per target continui.** I valori previsti degli insiemi per i target continui possono essere combinati utilizzando la media o la mediana dei valori previsti ricavati dai modelli di base.

Si noti che quando l'obiettivo è di ottimizzare la precisione del modello, le selezioni delle regole di combinazione vengono ignorate. Nel boosting viene sempre utilizzato un voto di maggiore ponderazione per il calcolo del punteggio degli obiettivi categoriali e una mediana pesata per il calcolo del punteggio dei target continui.

Boosting e bagging. Specificare il numero dei modelli di base da creare quando l'obiettivo è di ottimizzare la precisione o la stabilità del modello. Per il bagging, si tratta del numero di campioni di bootstrap. Questo valore deve essere un numero intero positivo.

Opzioni avanzate

Replica risultati. L'impostazione di un seme random consente di replicare le analisi. Il generatore di numeri random viene utilizzato per scegliere i record presenti nell'insieme di prevenzione del sovradattamento. Specificare un intero o fare clic su **Genera** per creare un intero pseudocasuale compreso tra 1 e 2147483647 incluso. Il valore predefinito è 54752075.

Opzioni del modello

Nome modello. È possibile generare il nome del modello automaticamente in base ai campi obiettivo oppure specificare un nome personalizzato. Il nome generato automaticamente è il nome del campo obiettivo.

Si noti che il valore previsto viene sempre calcolato quando viene eseguito il calcolo del punteggio del modello. Il nome del nuovo campo corrisponde al nome del campo obiettivo con l'aggiunta del prefisso *$L-*. Per esempio, se il campo obiettivo è denominato *vendite*, il nuovo campo si chiamerà *$L-vendite*.

Riepilogo del modello

La visualizzazione Riepilogo del modello è una snapshot, un riepilogo del modello e del suo adattamento.

Tabella. La tabella identifica alcune impostazioni del modello di alto livello, tra cui:

- Il nome dell'obiettivo specificato nella scheda Campi,
- L'esecuzione o meno della preparazione automatica dei dati come specificato nelle impostazioni Di base,
- Il metodo di selezione del modello ed i criteri di selezione specificati nelle impostazioni Selezione modello. Viene visualizzato anche il valore del criterio di selezione per il modello finale, presentato in un formato più piccolo e più pratico.

Grafico. Il grafico visualizza la precisione del modello finale, visualizzato in formato ingrandito. Il valore è 100 × il valore R^2 regolato per il modello finale.

Preparazione automatica dati

Questa visualizzazione contiene informazioni sui campi esclusi e sulla modalità di derivazione dei campi trasformati nel passaggio di preparazione automatica dei dati (ADP). Per ogni campo trasformato o escluso, la tabella indica il nome del campo, il ruolo nell'analisi e l'azione intrapresa nel passaggio dell'ADP. I campi sono ordinati in ordine alfabetico crescente in base al nome. Le possibili azioni intraprese per ogni campo sono:

- **Deriva durata: mesi** calcola il tempo trascorso in mesi dai valori in un campo che contiene date rispetto alla data di sistema corrente.
- **Deriva durata: ore** calcola il tempo trascorso in ore dai valori in un campo che contiene le ore rispetto all'ora di sistema corrente.
- **Trasforma livello di misurazione da continuo a ordinale** riformula i campi continui con meno di 5 valori univoci in campi ordinali.
- **Trasforma livello di misurazione da ordinale a continuo** riformula i campi ordinali con più di 10 valori univoci in campi continui.
- **Ritaglia valori anomali** imposta i valori dei predittori continui che si posizionano oltre un valore di interruzione (3 deviazioni standard dalla media) sul valore di interruzione.
- **Sostituisci i valori mancanti** sostituisce i valori mancanti dei campi nominali con la moda, dei campi ordinali con la mediana e dei campi continui con la media.
- **Unisci categorie poco dense per aumentare al massimo l'associazione all'obiettivo** identifica le categorie di predittori "simili" in base alla relazione tra input e obiettivo. Le categorie che non presentano differenze significative (ovvero che hanno un valore *p* superiore a 0,05) vengono unite.
- **Escludi predittore costante / dopo la gestione dei valori anomali / dopo l'unione delle categorie** elimina i predittori con un solo valore, possibilmente dopo che sono state completate le altre azioni ADP.

Importanza predittore

Generalmente, lo sforzo della modellazione viene concentrato sui campi predittore più importanti, senza considerare o ignorando i campi di minore importanza. Il grafico dell'importanza dei predittori rende più semplice questa operazione, indicando l'importanza relativa di ciascun predittore nella stima del modello. Poiché i valori sono relativi, la somma dei valori visualizzata per tutti i predittori è 1.0. L'importanza dei predittori non è correlata alla precisione del modello. Riguarda unicamente l'importanza di ciascun predittore nell'esecuzione di una previsione e non il grado di precisione della previsione.

Previsioni e osservazioni

Visualizza un grafico a dispersione in bin dei valori previsti sull'asse verticale in base ai valori osservati sull'asse orizzontale. In teoria i punti dovrebbero trovarsi su una linea a 45 gradi; da questa visualizzazione si può capire se il modello è particolarmente carente nella previsione di determinati record.

Residui

Visualizza un grafico diagnostico dei residui del modello.

Stili del grafico. Sono disponibili diversi stili di visualizzazione, a cui è possibile accedere dall'elenco a discesa **Stile**.

- **Istogramma.** È un istogramma in bin dei residui studentizzati con una sovrapposizione della distribuzione normale. I modelli lineari presumono che i residui abbiano una distribuzione normale, quindi l'istogramma idealmente dovrebbe avvicinarsi molto alla linea uniforme.
- **Grafico P-P.** È un grafico probabilità-probabilità in bin che confronta i residui studentizzati con una distribuzione normale. Se la pendenza del tracciato dei punti è minore rispetto alla linea normale, i residui mostrano una maggiore variabilità rispetto a una distribuzione normale. Se la pendenza è maggiore, i residui risultano meno variabili rispetto a una distribuzione normale. Se i punti del tracciato formano una curva a S, la distribuzione dei residui è asimmetrica.

Valori anomali

Questa tabella contiene i record che influenzano in modo anomalo il modello e visualizza ID dei record, se specificato nella scheda Campi, valore dell'obiettivo e distanza di Cook. La distanza di Cook è una misura di quanto cambierebbero i residui di tutti i record se un particolare record fosse escluso dal calcolo dei coefficienti del modello. Un valore elevato per la distanza di Cook indica che l'esclusione di un record modifica i coefficienti in modo sostanziale e deve quindi essere considerata come fattore influente.

I record influenti devono essere esaminati con attenzione per determinare se è possibile dare agli stessi meno peso nella stima del modello, troncare i valori anomali in corrispondenza di una soglia accettabile o eliminare completamente i record influenti.

Effetti

Questa visualizzazione mostra le dimensioni di ogni effetto nel modello.

Stili. Sono disponibili diversi stili di visualizzazione, a cui è possibile accedere dall'elenco a discesa **Stile**.

- **Diagramma.** In questo grafico, gli effetti sono ordinati dall'alto verso il basso per importanza predittore decrescente. Le linee di collegamento del diagramma vengono pesate in base alla significatività degli effetti, con la maggiore ampiezza della linea corrispondente agli effetti più significativi (minori valori *p*). Passando il mouse sopra una linea di collegamento si visualizza un testo di suggerimento che mostra il valore *p* e l'importanza dell'effetto. Questa è l'opzione di default.
- **Tabella.** Tabella ANOVA per gli effetti generali e specifici del modello. Gli effetti specifici sono ordinati dall'alto verso il basso in ordine decrescente in base all'importanza dei predittori. Si noti che, per default, la tabella viene compressa in modo da visualizzare solo i risultati del modello complessivo. Per visualizzare i risultati dei singoli effetti del modello, fare clic sulla cella **Modello corretto** all'interno della tabella.

Importanza predittore. È disponibile un dispositivo di scorrimento Importanza predittore che controlla i predittori mostrati nella visualizzazione. L'azione del dispositivo di scorrimento non modifica il modello, consente semplicemente di evidenziare i predittori più importanti. Per default sono visualizzati i primi 10 effetti.

Significatività. È disponibile un dispositivo di scorrimento Significatività che controlla ulteriormente gli effetti mostrati nella visualizzazione, oltre a quelli mostrati in base all'importanza predittore. Gli effetti con valori di significatività superiori al valore del dispositivo di scorrimento vengono nascosti. L'azione del dispositivo di scorrimento non modifica il modello, consente semplicemente di evidenziare gli effetti più importanti. Il valore di default è 1,00, ovvero gli effetti non vengono filtrati in base alla significatività.

Coefficienti

Questa visualizzazione mostra il valore di ogni coefficiente nel modello. Si noti che i fattori (predittori categoriali) sono codificati mediante un indicatore nel modello, in modo tale che agli **effetti** contenenti fattori possano essere associati più **coefficienti**, uno per ogni categoria esclusa la categoria corrispondente al parametro ridondante (di riferimento).

Stili. Sono disponibili diversi stili di visualizzazione, a cui è possibile accedere dall'elenco a discesa **Stile**.

- **Diagramma.** Questo grafico visualizza prima l'intercettazione, quindi ordina gli effetti dall'alto verso il basso per importanza predittore decrescente. Negli effetti che contengono fattori i coefficienti vengono ordinati in ordine crescente in base ai valori dei dati. Le linee di collegamento del diagramma sono colorate in base al segno del coefficiente (vedere la legenda del diagramma) e pesate in base alla significatività del coefficiente; la maggiore ampiezza della linea corrisponde ai coefficienti più significativi (valori *p* inferiori). Se si passa il mouse sopra una linea di collegamento, compare un testo di suggerimento che mostra il valore del coefficiente, il suo valore *p* e l'importanza dell'effetto a cui è associato il parametro. Questo è lo stile di default.
- **Tabella.** Indica i valori, i test di significatività e gli intervalli di confidenza per i singoli coefficienti del modello. Dopo l'intercettazione, gli effetti vengono ordinati dall'alto verso il basso in ordine decrescente in base all'importanza dei predittori. Negli effetti che contengono fattori i coefficienti vengono ordinati in ordine crescente in base ai valori dei dati. Si noti che, per default. la tabella è compressa in modo da mostrare solo il coefficiente, la significatività e l'importanza dei singoli parametri del modello. Per visualizzare l'errore standard, la statistica *t* e l'intervallo di confidenza, fare clic nella cella **Coefficiente** all'interno della tabella. Se si passa il mouse sopra il nome di un parametro del modello all'interno della tabella, viene visualizzato un testo di suggerimento che mostra il nome del parametro, l'effetto a cui questo è associato e, per i predittori categoriali, le etichette dei valori associate al parametro del modello. Questo può essere utile soprattutto per vedere le nuove categorie create quando la preparazione automatica dei dati unisce le categorie simili di un predittore categoriale.

Importanza predittore. È disponibile un dispositivo di scorrimento Importanza predittore che controlla i predittori mostrati nella visualizzazione. L'azione del dispositivo di scorrimento non modifica il modello, consente semplicemente di evidenziare i predittori più importanti. Per default sono visualizzati i primi 10 effetti.

Significatività. È disponibile un dispositivo di scorrimento Significatività che controlla ulteriormente i coefficienti mostrati nella vista, oltre a quelli mostrati in base all'importanza predittore. I coefficienti con valori di significatività superiori al valore del dispositivo di scorrimento vengono nascosti. L'azione del dispositivo di scorrimento non modifica il modello, consente semplicemente di evidenziare i coefficienti più importanti. Il valore di default è 1,00, ovvero i coefficienti non vengono filtrati in base alla significatività.

Medie stimate

Vengono visualizzati grafici relativi ai predittori significativi. Ogni grafico visualizza il valore del modello stimato relativo all'obiettivo sull'asse verticale per ogni valore del predittore sull'asse orizzontale, mantenendo costanti tutti gli altri predittori. È una visualizzazione utile degli effetti dei coefficienti di ciascun predittore sull'obiettivo.

Nota: se non sono presenti predittori significativi, non viene prodotta alcuna media stimata.

Riepilogo di creazione del modello

Quando si sceglie un algoritmo di selezione del modello diverso da **Nessuno** in Selezione modello, vengono visualizzati alcuni dettagli del processo di creazione del modello.

Stepwise in avanti. Quando l'algoritmo di selezione è Stepwise in avanti, la tabella visualizza le ultime 10 fasi nell'algoritmo stepwise. Per ogni passaggio, vengono visualizzati il valore del criterio di selezione e gli effetti nel modello in corrispondenza di tale passaggio. Ciò consente di verificare l'entità del contributo di ogni passaggio al modello. Ciascuna colonna consente di ordinare le righe in modo tale da individuare più facilmente gli effetti presenti nel modello in un determinato passaggio.

Sottoinsiemi migliori. Quando l'algoritmo di selezione è Sottoinsiemi migliori, la tabella visualizza i primi 10 modelli. Per ogni modello, vengono visualizzati il valore del criterio di selezione e gli effetti presenti nel modello. Ciò consente di verificare la stabilità dei modelli più importanti. Se i modelli tendono ad avere molti effetti simili con poche differenze, il modello migliore può essere considerato affidabile, se invece i modelli contengono effetti molto diversi, alcuni effetti potrebbero essere troppo

simili e sarebbe opportuno combinarli (o eliminarne uno). Ciascuna colonna consente di ordinare le righe in modo tale da individuare più facilmente gli effetti presenti nel modello in un determinato passaggio.

Impostazioni

Si noti che il valore previsto viene sempre calcolato quando viene eseguito il calcolo del punteggio del modello. Il nome del nuovo campo corrisponde al nome del campo obiettivo con l'aggiunta del prefisso *$L-*. Per esempio, se il campo obiettivo è denominato *vendite*, il nuovo campo si chiamerà *$L-vendite*.

Genera SQL per questo modello. Quando si utilizzano dati da un database, è possibile "restituire" codice SQL al database per l'esecuzione, migliorando in tal modo le performance di molte operazioni.

Calcola il punteggio convertendo in SQL nativo. Se questa opzione è selezionata, genera il codice SQL per calcolare in modo nativo il punteggio del modello all'interno dell'applicazione.

Nodo Logistica

La regressione logistica, conosciuta anche come **regressione nominale**, è una tecnica statistica per classificare i record in base ai valori dei campi di input. È analoga alla regressione lineare ma, al posto di un campo numerico, prende un campo obiettivo categoriale. Sono supportati sia i modelli binomiali (per gli obiettivi con due categorie discrete), sia quelli multinomiali (per gli obiettivi con più di due categorie).

La regressione logistica consente di creare un insieme di equazioni che correlano i valori dei campi di input alle probabilità associate a ciascuna categoria del campo di output. Dopo essere stato generato, il modello può essere utilizzato per stimare le probabilità per i nuovi dati. Per ciascun record, viene calcolata la probabilità di appartenenza per ciascuna categoria di output possibile. La categoria obiettivo con la maggiore probabilità viene assegnata come valore di output previsto per quel record.

Esempio binomiale. Una società di telecomunicazioni è preoccupata per il numero di clienti che passano alla concorrenza. Con i dati relativi all'utilizzo del servizio è possibile creare un modello binomiale per prevedere quali clienti sono inclini a passare a un altro operatore e personalizzare le offerte in modo da conservare il maggior numero possibile di clienti. Viene utilizzato un modello binomiale perché il target dispone di due categorie distinte (con probabilità o meno di eseguire il passaggio alla concorrenza).

Nota: solo per i modelli binomiali, i campi stringa devono essere limitati a 8 caratteri. Se necessario, le stringhe più lunghe possono essere ricodificate utilizzando un nodo Ricodifica.

Esempio multinomiale. Un provider di telecomunicazioni ha segmentato la base clienti per modelli di utilizzo del servizio, suddividendo i clienti in quattro categorie. Utilizzando i dati demografici per prevedere l'appartenenza al gruppo, è possibile creare un modello multinomiale per classificare i potenziali clienti in gruppi e personalizzare le offerte per i singoli clienti.

Requisiti. Uno o più campi di input ed esattamente un campo obiettivo categoriale con due o più categorie. Per un modello binomiale, l'obiettivo deve avere un livello di misurazione *Flag*. Per un modello multinomiale, l'obiettivo può avere un livello di misurazione *Flag* o *Nominale* con almeno due categorie. I campi impostati su *Entrambe* o *Nessuna* verranno ignorati. È necessario che i tipi dei campi utilizzati nel modello siano completamente istanziati.

Efficacia. I modelli di regressione logistica spesso sono molto precisi. Sono in grado di gestire campi di input sia simbolici sia numerici e possono fornire le probabilità previste per tutte le categorie obiettivo, consentendo così di identificare facilmente la seconda migliore ipotesi. I modelli logistici sono particolarmente efficaci quando l'appartenenza a un gruppo è un campo veramente categoriale; se l'appartenenza a un gruppo si basa sui valori di un campo intervallo continuo (per esempio QI alto e QI basso), è opportuno valutare l'impiego della regressione lineare per sfruttare la maggiore quantità di informazioni utili offerta dall'intervallo completo di valori. I modelli logistici possono anche eseguire la selezione automatica del campo, sebbene altri approcci, come, ad esempio, i modelli di struttura ad albero o la selezione funzioni, possano effettuare tale operazione più rapidamente sui dataset di grandi

dimensioni. Infine, poiché sono di facile comprensione per molti analisti e data miner, i modelli logistici possono essere utilizzati come riferimento con cui confrontare altre tecniche di modellazione.

Quando si elaborano insiemi di dati di grandi dimensioni, per migliorare notevolmente le prestazioni si consiglia di disattivare l'opzione di output avanzato Test rapporto di verosimiglianza. Per ulteriori informazioni, consultare l'argomento "Output delle opzioni avanzate della regressione logistica" a pagina 174.

Opzioni del modello di nodo Logistica

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.

Utilizza dati partizionati. Se è definito un campo partizione, questa opzione garantisce che per la creazione del modello vengano utilizzati dati solo dalla partizione di addestramento.

Crea modelli di suddivisione. Crea un modello separato per ogni valore possibile dei campi di input che sono stati specificati come campi di suddivisione. Per ulteriori informazioni, consultare l'argomento "Creazione di modelli di suddivisione" a pagina 28.

Procedura. Specifica se deve essere creato un modello binomiale o multinomiale. Le opzioni disponibili nella finestra di dialogo possono variare a seconda del tipo di procedura di modellazione selezionato.

- **Binomiale.** Utilizzato quando il campo obiettivo è di tipo flag o nominale con due valori discreti (dicotomico), quali *sì/no*, *attivato/disattivato*, *maschio/femmina*.
- **Multinomiale.** Utilizzato quando il campo obiettivo è un campo nominale con più di due valori. È possibile specificare **Effetti principali**, **Fattoriale completo** o **Personalizzato**.

Includi termine costante nell'equazione. Questa opzione determina se le equazioni risultanti includeranno un termine costante. Nella maggior parte dei casi, è opportuno lasciare selezionata questa opzione.

Modelli binomiali

Per i modelli binomiali sono disponibili i seguenti metodi e opzioni:

Metodo. Specificare il metodo da utilizzare per la generazione del modello di regressione logistica.

- **Invio.** Questo è il metodo di default che immette tutti i termini direttamente nell'equazione. Nella generazione del modello non viene eseguita alcuna selezione di campo.
- **In avanti.** Il metodo di selezione dei campi In avanti crea il modello spostandosi in avanti un passo alla volta. Con questo metodo, il modello iniziale è quello più semplice e solo la costante e i termini possono essere aggiunti al modello. Ad ogni passo successivo, i termini che non si trovano ancora nel modello vengono testati per valutare fino a che punto lo migliorano e il migliore di essi viene aggiunto al modello. Quando non è più possibile aggiungere altri termini oppure quando il miglior termine candidato non produce un miglioramento sufficiente del modello, viene generato il modello finale.
- **All'indietro.** Il metodo All'indietro essenzialmente è l'opposto del metodo In avanti. Con questo metodo, il modello iniziale contiene tutti i termini come predittori ed è possibile solo rimuoverli dal modello. I termini modello che contribuiscono poco alla generazione del modello vengono rimossi uno per uno, finché non è più possibile rimuoverne altri senza peggiorare notevolmente il modello. Il modello finale viene generato con i termini restanti.

Input categoriali. Elenca i campi identificati come categoriali, ovvero quelli con livello di misurazione flag, nominale o ordinale. Per ogni campo categoriale è possibile specificare il contrasto e la categoria di base.

- **Nome campo.** Questa colonna contiene i nomi dei campi degli input categoriali ed è precompilata con tutti i valori dei flag e nominali presenti nei dati. Per aggiungere input continui o numerici a questa colonna, fare clic sull'icona di aggiunta dei campi a destra dell'elenco e selezionare gli input desiderati.
- **Contrasto.** L'interpretazione dei coefficienti di regressione per un campo categoriale dipende dai contrasti utilizzati. Il contrasto determina la modalità di impostazione dei test di ipotesi per il confronto delle medie stimate. Per esempio, se è noto che un campo categoriale ha un ordine implicito quale uno schema o un raggruppamento, è possibile utilizzare il contrasto per modellare quell'ordine. I contrasti disponibili sono:

  Indicatore. I contrasti indicano l'appartenenza o la non appartenenza alla categoria. È il metodo predefinito.

  Semplice. Ogni categoria del campo predittore, tranne quella di riferimento, viene confrontata con la categoria di riferimento.

  Differenza. Ogni categoria del campo predittore, tranne la prima, viene confrontata con l'effetto medio delle categorie precedenti. Sono noti anche come contrasti inversi di Helmert.

  Helmert. Ogni categoria del campo predittore, tranne l'ultima, viene confrontata con l'effetto medio delle categorie successive.

  Ripetuto. Ogni categoria del campo predittore, tranne la prima, viene confrontata con la categoria che la precede.

  Polinomiale. Contrasti polinomiali ortogonali. Si presume che le categorie siano equamente distanziate. I contrasti polinomiali sono disponibili solo per i campi numerici.

  Deviazione. Ogni categoria del campo predittore, tranne quella di riferimento, viene confrontata con l'effetto globale.
- **Categoria di base.** Specifica come viene determinata la categoria di riferimento per il tipo di contrasto selezionato. Selezionare **Prima** per utilizzare la prima categoria per il campo di input —in ordine alfabetico— oppure selezionare **Ultima** per utilizzare l'ultima categoria. L'impostazione di default è Prima.

  Nota: questo campo non è disponibile se il contrasto è impostato su Differenza, Helmert, Ripetuto o Polinomiale.

La stima dell'effetto di ogni campo sulla risposta globale viene calcolata come aumento o diminuzione nella verosimiglianza di ciascuna delle altre categorie relative alla categoria di riferimento. Questo può facilitare l'individuazione dei campi e dei valori che hanno le maggiori probabilità di fornire una risposta specifica.

La categoria di base è visualizzata nell'output come 0.0, poiché il confronto con se stessa genera un risultato vuoto. Tutte le altre categorie sono visualizzate come equazioni inerenti alla categoria di base. Per ulteriori informazioni, consultare l'argomento "Dettagli del nugget del modello Logistica" a pagina 177.

Modelli multinomiali

Per i modelli multinomiali sono disponibili i seguenti metodi e opzioni:

Metodo. Specificare il metodo da utilizzare per la generazione del modello di regressione logistica.

- **Invio.** Questo è il metodo di default che immette tutti i termini direttamente nell'equazione. Nella generazione del modello non viene eseguita alcuna selezione di campo.
- **Stepwise.** Il metodo Stepwise di selezione dei campi consente di creare l'equazione passo per passo. Il modello iniziale è il più semplice possibile e l'equazione non presenta termini di modello (ad eccezione della costante). A ogni fase vengono valutati i termini che ancora non sono stati aggiunti al modello e, se il migliore di essi aumenta notevolmente il potere predittivo del modello, viene aggiunto. Inoltre, i termini che al momento si trovano nel modello vengono rivalutati per stabilire se è possibile rimuoverli senza pregiudicare eccessivamente il modello. Se questa possibilità esiste, i campi vengono rimossi. Il

processo viene ripetuto e vengono aggiunti e/o rimossi altri termini. Quando non è più possibile aggiungere altri termini per migliorare il modello e non è più possibile rimuoverne senza danneggiare il modello, viene generato il modello finale.

- **In avanti.** Il metodo di selezione dei campi In avanti è simile al metodo Stepwise, poiché anche in questo caso il modello viene generato passo per passo. Tuttavia, con questo metodo, il modello iniziale è quello più semplice e solo la costante e i termini possono solo essere aggiunti al modello. Ad ogni passo successivo, i termini che non si trovano ancora nel modello vengono testati per valutare fino a che punto lo migliorano e il migliore di essi viene aggiunto al modello. Quando non è più possibile aggiungere altri termini oppure quando il miglior termine candidato non produce un miglioramento sufficiente del modello, viene generato il modello finale.
- **All'indietro.** Il metodo All'indietro essenzialmente è l'opposto del metodo In avanti. Con questo metodo, il modello iniziale contiene tutti i termini come predittori ed è possibile solo rimuoverli dal modello. I termini modello che contribuiscono poco alla generazione del modello vengono rimossi uno per uno, finché non è più possibile rimuoverne altri senza peggiorare notevolmente il modello. Il modello finale viene generato con i termini restanti.
- **Stepwise all'indietro.** Il metodo Stepwise All'indietro essenzialmente è l'opposto del metodo Stepwise. Con questo metodo, il modello iniziale contiene tutti i termini come predittori. Ad ogni passo, i termini nel modello vengono valutati e tutti quelli che possono essere rimossi senza pregiudicare il modello vengono eliminati. Inoltre, i termini precedentemente rimossi vengono rivalutati per stabilire se il migliore di essi aumenta notevolmente il potere predittivo del modello. In tal caso, viene nuovamente aggiunto al modello. Quando non è più possibile rimuovere altri termini per migliorare il modello e non è più possibile aggiungerne senza danneggiarlo notevolmente, viene generato il modello finale.

Nota: i metodi automatici (compresi i metodi Stepwise, In avanti e All'indietro) sono metodi di apprendimento di elevata adattabilità e hanno una forte tendenza a sovradattare i dati di addestramento. Quando vengono utilizzati questi metodi, è particolarmente importante verificare la validità del modello risultante con dati nuovi o con un campione estratto per il test mediante il nodo Partizione.

Categoria di base per obiettivo. Specifica come viene determinata la categoria di riferimento. Essa viene utilizzata come riferimento rispetto a cui vengono stimate le equazioni di regressione per tutte le altre categorie dell'obiettivo. Selezionare **Prima** per utilizzare la prima categoria per il campo obiettivo corrente —in ordine alfabetico— oppure selezionare **Ultima** per utilizzare l'ultima categoria. In alternativa, è possibile selezionare **Specifica** per scegliere una categoria specifica e selezionare il valore desiderato dall'elenco. È possibile definire i valori disponibili per ogni campo in un nodo Tipo.

Spesso come categoria di base viene specificata la categoria di minore interesse, per esempio un prodotto civetta. Le altre categorie vengono quindi correlate a questa categoria di base in maniera relativa, per individuare in base a quale elemento è più probabile che rientrino in una categoria separata. Questo può facilitare l'individuazione dei campi e dei valori che hanno le maggiori probabilità di fornire una risposta specifica.

La categoria di base è visualizzata nell'output come 0.0, poiché il confronto con se stessa genera un risultato vuoto. Tutte le altre categorie sono visualizzate come equazioni inerenti alla categoria di base. Per ulteriori informazioni, consultare l'argomento “Dettagli del nugget del modello Logistica” a pagina 177.

Tipo di modello. Esistono tre opzioni per definire i termini nel modello. I modelli **Effetti principali** includono solo i campi di input individualmente e non verificano le interazioni (effetti di moltiplicazione) tra i vari campi di input. I modelli **Fattoriale completo** includono tutte le interazioni e gli effetti principali dei campi di input. I modelli Fattoriale completo riescono a catturare meglio le relazioni complesse ma sono anche di più difficile interpretazione e più soggetti al sovradattamento. A causa del numero di combinazioni potenzialmente elevato, i metodi di selezione automatica dei campi (diversi da Per blocchi) vengono disattivati per i modelli Fattoriale completo. I modelli **Personalizzato** includono solo i termini (effetti e iterazioni principali) specificati dall'utente. Quando si seleziona questa opzione, utilizzare l'elenco Termini di modello per aggiungere o rimuovere termini nel modello.

Termini di modello. Quando si crea un modello Personalizzato, è necessario specificare esplicitamente i termini nel modello. Nell'elenco è riportato l'insieme corrente di termini per il modello. I pulsanti sul lato destro dell'elenco Termini di modello consentono di aggiungere e rimuovere i termini di modello.

- Per aggiungere termini al modello, fare clic sul pulsante *Aggiunge nuovi termini di modello*.
- Per eliminare termini, selezionare quelli desiderati e fare clic sul pulsante *Elimina i termini di modello selezionati*.

Aggiunta di termini a un modello di regressione logistica

Quando si richiede un modello di regressione logistica personalizzato, è possibile aggiungere termini al modello facendo clic sul pulsante *Aggiunge nuovi termini di modello* nella scheda del modello di regressione logistica. Viene visualizzata la finestra di dialogo Nuovi termini in cui è possibile specificare i termini.

Tipo di termine da aggiungere. È possibile aggiungere termini al modello in vari modi, in relazione alla selezione dei campi di input nell'elenco Campi disponibili.

- **Interazione singola.** Viene inserito il termine che rappresenta l'interazione di tutti i campi selezionati.
- **Effetti principali.** Viene inserito un termine di effetto principale (il campo stesso) per ogni campo di input selezionato.
- **Tutte le interazioni a 2 vie.** Viene inserito un termine di interazione a due vie (il prodotto dei campi di input) per ogni coppia possibile di campi di input selezionata. Per esempio, se sono stati selezionati i campi di input *A*, *B* e *C* nell'elenco Campi disponibili, questo metodo inserirà i termini *A * B*, *A * C* e *B * C*.
- **Tutte le interazioni a 3 vie.** Viene inserito un termine di interazione a tre vie (il prodotto dei campi di input) per ogni combinazione possibile di campi di input selezionata, prendendone in considerazione tre per volta. Per esempio, se sono stati selezionati i campi di input *A*, *B*, *C* e *D* nell'elenco Campi disponibili, questo metodo inserirà i termini *A * B * C*, *A * B * D*, *A * C * D* e *B * C * D*.
- **Tutte le interazioni a 4 vie.** Viene inserito un termine di interazione a quattro vie (il prodotto dei campi di input) per ogni combinazione possibile di campi di input selezionata, prendendone in considerazione quattro per volta. Per esempio, se sono stati selezionati i campi di input *A*, *B*, *C*, *D* e *E* nell'elenco Campi disponibili, questo metodo inserirà i termini *A * B * C * D*, *A * B * C * E*, *A * B * D * E*, *A * C * D * E* e *B * C * D * E*.

Campi disponibili. Elenca i campi di input disponibili da utilizzare nella creazione dei termini di modello.

Anteprima. Mostra i termini che verranno aggiunti al modello se si seleziona **Inserisci**, in base ai campi selezionati e al tipo di termine.

Inserisci. Inserisce i termini nel modello (in base alla selezione corrente di campi e del tipo di termine) e chiude la finestra di dialogo.

Opzioni della scheda Livello avanzato del nodo Logistica

Le opzioni avanzate consentono agli utenti esperti della regressione logistica di ottimizzare il processo di addestramento. Per accedere alle opzioni avanzate, impostare la modalità su **Livello avanzato** nella scheda Livello avanzato.

Scala (solo modelli multinomiali). È possibile specificare un valore di graduazione della dispersione che verrà utilizzato per correggere la stima della matrice di covarianza dei parametri. **Pearson** stima il valore di graduazione utilizzando la statistica chi-quadrato Pearson. **Devianza** stima il valore di graduazione utilizzando la statistica della funzione di devianza (chi-quadrato del rapporto di verosimiglianza). È anche possibile specificare un valore di graduazione definito dall'utente. che deve essere un valore numerico positivo.

Accoda tutte le probabilità. Se questa opzione è selezionata, a ogni record elaborato dal nodo verranno aggiunte le probabilità per ciascuna categoria del campo di output. Se questa opzione non è selezionata, verrà aggiunta solo la probabilità della categoria prevista.

Per esempio, una tabella contenente i risultati di un modello multinomiale con tre categorie comprenderà cinque nuove colonne. In una colonna sarà riportata la probabilità che il risultato venga previsto correttamente, in quella seguente la probabilità che questa previsione sia corretta o errata e in altre tre colonne sarà visualizzata la probabilità che la previsione di ogni categoria sia errata o corretta. Per ulteriori informazioni, consultare l'argomento "Nugget del modello Logistica" a pagina 176.

Nota: questa opzione è sempre selezionata per i modelli binomiali.

Tolleranza della singolarità. Specificare la tolleranza utilizzata nel controllo delle singolarità.

Convergenza. Queste opzioni consentono di controllare i parametri per la convergenza del modello. Quando viene eseguito il modello, le impostazioni per la convergenza controllano quante volte vengono eseguiti ripetutamente i vari parametri per verificarne l'adeguatezza dell'adattamento. Quanto maggiore è la frequenza con cui vengono eseguiti i parametri, tanto più vicini saranno i risultati (ovvero, i risultati convergeranno). Per ulteriori informazioni, consultare l'argomento "Opzioni di convergenza di Regressione logistica".

Output. Queste opzioni consentono di richiedere ulteriori statistiche che verranno visualizzate nell'output avanzato del nugget del modello creato dal nodo. Per ulteriori informazioni, consultare l'argomento "Output delle opzioni avanzate della regressione logistica".

Controllo. Queste opzioni consentono di controllare i criteri per l'aggiunta e la rimozione di campi con i metodi di stima Stepwise, In avanti, All'indietro o Stepwise all'indietro. Se il metodo selezionato è Per blocchi, il pulsante è disattivato. Per ulteriori informazioni, consultare l'argomento "Opzioni di controllo del nodo Regressione logistica" a pagina 175.

Opzioni di convergenza di Regressione logistica

È possibile impostare i parametri di convergenza per la stima del modello di regressione logistica.

Numero massimo di iterazioni. Specificare il numero massimo di iterazioni per la stima del modello.

Massimo numero di dimezzamenti. Il dimezzamento è una tecnica utilizzata dalla regressione logistica per affrontare le complessità nel processo di stima. In circostanze normali, è necessario utilizzare l'impostazione di default.

Convergenza verosimiglianza logaritmica. Le iterazioni si interrompono se la variazione relativa della verosimiglianza è inferiore a questo valore. Il criterio non viene utilizzato se il valore specificato è 0.

Convergenza parametri. Le iterazioni si interrompono se la variazione assoluta o relativa delle stime dei parametri è inferiore a questo valore. Il criterio non viene utilizzato se il valore specificato è 0.

Delta (solo modelli multinomiali). È possibile specificare un valore tra 0 e 1 da aggiungere a ciascuna cella vuota (combinazione di valori dei campi di input e di output). Questa impostazione può aiutare l'algoritmo di stima a gestire i dati nel caso in cui vi sia un numero eccessivo di combinazioni possibili di valori di campo rispetto al numero di record nei dati. Il valore predefinito è 0.

Output delle opzioni avanzate della regressione logistica

Selezionare l'output facoltativo che si desidera visualizzare nell'output avanzato del nugget del modello Regressione. Per visualizzare l'output avanzato, individuare il nugget del modello e fare clic sulla scheda **Opzioni avanzate**. Per ulteriori informazioni, consultare l'argomento "Output avanzato del nugget del modello Logistica" a pagina 178.

Opzioni binomiali

Selezionare i tipi di output da generare per il modello. Per ulteriori informazioni, consultare l'argomento “Output avanzato del nugget del modello Logistica” a pagina 178.

Visualizzazione. Selezionare se visualizzare i risultati a ogni passaggio o attendere il completamento di tutti i passaggi.

CI per exp(B). Selezionare gli intervalli di confidenza per ogni coefficiente (visualizzato come Beta) dell'espressione. Specificare il livello dell'intervallo di confidenza (l'impostazione di default è 95%).

Diagnosi dei residui. Richiede una tabella Diagnostiche per casi dei residui.

- **Valori anomali all'esterno (dev. std.).** Elenca solo i casi dei residui per cui il valore standardizzato assoluto della variabile riportata è grande almeno quanto il valore specificato. Il valore predefinito è 2.
- **Tutti i casi.** Include tutti i casi nella tabella Diagnostiche per casi dei residui.

  Nota: poiché questa opzione riporta ogni singolo record di input, la tabella generata nel report può risultare eccezionalmente grande, con una riga per ogni record.

Valore di interruzione della classificazione. Consente di determinare il punto di divisione per la classificazione dei casi. I casi con valori attesi che superano il valore di riferimento sono classificati come positivi, mentre i casi con valori previsti minori del valore di riferimento sono classificati come negativi. Per modificare il valore predefinito, inserire un valore compreso tra 0,01 e 0,99.

Opzioni multinomiali

Selezionare i tipi di output da generare per il modello. Per ulteriori informazioni, consultare l'argomento “Output avanzato del nugget del modello Logistica” a pagina 178.

Nota: la selezione dell'opzione **Test rapporto di verosimiglianza** aumenta notevolmente il tempo di elaborazione richiesto per generare un modello di regressione logistica. Se la creazione del modello richiede troppo tempo, è possibile disattivare questa opzione o utilizzare le statistiche Wald e Punteggio. Per ulteriori informazioni, consultare l'argomento “Opzioni di controllo del nodo Regressione logistica”.

Cronologia iterazione ogni. Selezionare l'intervallo tra i passaggi per stampare lo stato dell'iterazione nell'output avanzato.

Intervallo di confidenza. Gli intervalli di confidenza per i coefficienti nelle equazioni. Specificare il livello dell'intervallo di confidenza (l'impostazione di default è 95%).

Opzioni di controllo del nodo Regressione logistica

Queste opzioni consentono di controllare i criteri per l'aggiunta e la rimozione di campi con i metodi di stima Stepwise, In avanti, All'indietro o Stepwise all'indietro.

Numero di termini nel modello (solo modelli multinomiali). È possibile specificare il numero minimo di termini nel modello per i modelli All'indietro e Stepwise All'indietro e il numero massimo di termini per i modelli In avanti e Stepwise. Se si specifica un valore minimo superiore a 0, il modello includerà molti termini, anche se alcuni termini sarebbero stati rimossi in base a criteri statistici. L'impostazione minima viene ignorata per i modelli In avanti, Stepwise e Per blocchi. Se si specifica un valore massimo, alcuni termini potrebbero essere omessi nel modello, anche se sarebbero stati selezionati in base a un criterio statistico. L'impostazione **Massimo** viene ignorata per i modelli All'indietro, Stepwise All'indietro e Per blocchi.

Criterio di immissione (solo modelli multinomiali). Selezionare **Punteggio** per aumentare al massimo la velocità di elaborazione. L'opzione **Rapporto di verosimiglianza** può fornire stime più solide ma richiede tempi di elaborazione maggiori. Per default viene utilizzata la statistica Punteggio.

Criterio di rimozione. Selezionare **Rapporto di verosimiglianza** per ottenere un modello più solido. Per diminuire i tempi richiesti per la creazione del modello, è possibile selezionare **Wald**. Se tuttavia la separazione dei dati è completa o quasi completa (per determinarla è possibile utilizzare la scheda Opzioni avanzate nel nugget del modello), la statistica di Wald diventa inaffidabile e non è consigliabile utilizzarla. Per default viene utilizzata la statistica del rapporto di verosimiglianza. Per i modelli binomiali è disponibile l'opzione aggiuntiva **Condizionale** che consente di eseguire la verifica dell'eliminazione in base alla probabilità della statistica del rapporto di verosimiglianza basato sulle stime condizionali dei parametri.

Soglie di significatività per i criteri RL. Questa opzione consente di specificare i criteri di seleziona in base alla probabilità statistica (il valore *p*) associata a ciascun campo. I campi verranno aggiunti al modello solo se il valore *p* associato è inferiore al valore di **inserimento** e verranno rimossi solo se il valore *p* è maggiore del valore di **eliminazione**. Il valore di **inserimento** deve essere inferiore al valore di **eliminazione**.

Requisiti per l'immissione o per l'eliminazione (solo modelli multinomiali). Per alcune applicazioni, non è opportuno aggiungere termini di interazione al modello a meno che in esso non siano contenuti anche termini di ordine inferiore per i campi coinvolti nel termine di interazione. Per esempio, non è consigliabile includere *A * B* nel modello a meno che non siano inclusi anche *A* e *B* nel modello. Queste opzioni consentono di stabilire il modo in cui vengono gestite tali dipendenze durante la selezione dei termini stepwise.

- **Gerarchia per effetti discreti.** Gli effetti di ordine superiore (interazioni che coinvolgono più campi) verranno immessi nel modello solo se tutti gli effetti di ordine inferiore (effetti principali o interazioni che coinvolgono pochi campi) per i campi significativi sono già inclusi nel modello, e gli effetti di ordine inferiore non verranno rimossi se nel modello sono presenti effetti di ordine superiore che coinvolgono gli stessi campi. Questa opzione è applicabile solo ai campi categoriali.
- **Gerarchia per tutti gli effetti.** Questa opzione funziona come l'opzione precedente, ad eccezione del fatto che si applica a tutti i campi di input.
- **Contenimento per tutti gli effetti.** Gli effetti possono essere inclusi nel modello solo se sono inclusi nel modello anche tutti gli effetti contenuti nell'effetto. Questa opzione è simile a **Gerarchia per tutti gli effetti**, con l'eccezione che i campi continui vengono trattati diversamente. Affinché un effetto contenga un altro effetto è necessario che l'effetto contenuto (ordine inferiore) includa *tutti* i campi continui coinvolti nell'effetto contenente (ordine superiore) e i campi categoriali dell'effetto contenuto devono essere un sottoinsieme di quelli nell'effetto contenente. Per esempio, se *A* e *B* sono campi categoriali e *X* è un campo continuo, il termine *A * B * X* contiene i termini *A * X* e *B * X*.
- **Nessuna.** Non sono applicate relazioni. I termini vengono aggiunti e rimossi nel modello in modo indipendente.

Nugget del modello Logistica

Un nugget del modello Logistica rappresenta l'equazione stimata da un nodo Logistica. Contiene tutte le informazioni intercettate dal modello Regressione logistica, nonché informazioni sulle prestazioni e la struttura del modello. Questo tipo di equazione può inoltre essere generato da altri modelli quali Oracle SVM.

Quando viene eseguito un flusso che contiene un nugget del modello Logistica, il nodo aggiunge due nuovi campi contenenti la previsione del modello e la probabilità associata. I nomi dei nuovi campi derivano dal nome del campo di output di cui si sta eseguendo la previsione, a cui viene aggiunto il prefisso *$L-* per la categoria prevista e *$LP-* per la probabilità associata. Per esempio, per un campo di output denominato *colorepref*, i nuovi campi si chiameranno *$L-colorepref* e *$LP-colorepref*. Inoltre, se viene selezionata l'opzione **Accoda tutte le probabilità** nel nodo Logistica, per ogni categoria del campo di output verrà aggiunto un ulteriore campo contenente la probabilità relativa alla categoria corrispondente per ogni record. Questi campi aggiuntivi vengono denominati in base ai valori del campo di output, con l'aggiunta del prefisso *$LP-*. Per esempio, se i valori legali di *colorpref* sono *Rosso*, *Verde* e *Blu*, verranno aggiunti tre nuovi campi: *$LP-Rosso*, *$LP-Verde* e *$LP-Blu*.

Generazione di un nodo Filtro. Il menu Genera consente di creare un nuovo nodo Filtro per passare i campi di input in base ai risultati del modello. I campi non utilizzati dal modello o eliminati dallo stesso per via della multicollinearità verranno filtrati dal nodo generato.

Dettagli del nugget del modello Logistica

Per i modelli multinomiali, la scheda Modello in un nugget del modello Logistica presenta una visualizzazione ripartita con le equazioni del modello nel riquadro di sinistra e l'importanza dei predittori sulla destra. Per i modelli binomiali, la scheda visualizza solo l'importanza dei predittori. Per ulteriori informazioni, consultare l'argomento “Importanza predittore” a pagina 43.

Equazioni dei modelli

Per i modelli multinomiali, il riquadro di sinistra visualizza le equazioni effettive stimate per il modello di regressione logistica. Esiste solo un'equazione per ogni categoria nel campo obiettivo, con l'eccezione della categoria di riferimento. Le equazioni vengono visualizzate nel formato TREE. Questo tipo di equazione può inoltre essere generato da alcuni altri modelli quali Oracle SVM.

Equazione per. Mostra le equazioni di regressione utilizzate per derivare le probabilità della categoria di destinazione, dato un insieme di valori predittori. L'ultima categoria del campo di destinazione viene considerata la **categoria di riferimento**; le equazioni visualizzate contengono gli odd logaritmici per le altre categorie di destinazione relative alla categoria di riferimento per un determinato insieme di valori predittivi. La probabilità prevista per ogni categoria del modello predittivo dato deriva da questi valori odd logaritmici.

Per calcolare le probabilità

Ogni equazione calcola gli odd logaritmici di una determinata categoria di destinazione relativa alla categoria di riferimento. Gli **odd logaritmici**, anche detti **logit**, sono il rapporto della probabilità per la categoria di destinazione specificata verso quello della categoria di riferimento, con la funzione del logaritmo naturale applicata al risultato. Per la categoria di riferimento, il rapporto odds della categoria relativa a se stessa è 1.0, e quindi gli odd logaritmici sono 0. È possibile considerarlo un'equazione implicita per la categoria di riferimento in cui tutti i coefficienti sono uguali a zero.

Per derivare la probabilità dagli odd logaritmici di una determinata categoria di destinazione, prendere il valore logit calcolato dall'equazione per tale categoria e applicare la seguente formula:

$P(gruppo_{\,i}) = \exp(g_{\,i}) \;/\; \Sigma_k \exp(g_{\,k})$

dove g indica gli odd logaritmici calcolate, i è l'indice di categoria e k va da 1 al numero delle categorie di destinazione.

Importanza predittore

Facoltativamente, nella scheda Modello, è possibile visualizzare anche un grafico che indica l'importanza relativa di ogni predittore nella stima del modello. In genere è opportuno concentrare la modellazione sui predittori più importanti e valutare se eliminare o ignorare quelli meno importanti. Notare che tale grafico è disponibile solo se è selezionata l'opzione **Calcola importanza predittore** nella scheda Analizza prima di generare il modello. Per ulteriori informazioni, consultare l'argomento “Importanza predittore” a pagina 43.

Nota: per calcolare l'importanza predittore per la regressione logistica potrebbe essere necessario più tempo rispetto a quanto necessario per il calcolo per altri tipi di modelli; inoltre, l'importanza predittore non è selezionata nella scheda Analizza per impostazione predefinita. La selezione di questa opzione potrebbe rallentare le prestazioni, in particolare con insiemi di dati di grandi dimensioni.

Riepilogo del nugget del modello Logistica

Il riepilogo di un modello di regressione logistica visualizza i campi e le impostazioni utilizzati per generare il modello. Inoltre, se è stato eseguito un nodo Analisi collegato a questo nodo Modelli, in questa sezione verranno visualizzate anche le informazioni sui risultati di tale analisi. Per informazioni generali relative all'utilizzo del browser del modello, consultare “Esplorazione dei nugget del modello” a pagina 42.

Impostazioni del nugget del modello Logistica

La scheda Impostazioni di un nugget del modello Logistica specifica le opzioni per le confidenze, le probabilità, i punteggi di propensione e la generazione SQL durante il calcolo del punteggio del modello. Questa scheda è disponibile solo dopo che il nugget del modello è stato aggiunto a un flusso e visualizza diverse opzioni a seconda del tipo di modello e obiettivo.

Modelli multinomiali

Per i modelli multinomiali sono disponibili le seguenti opzioni:

Calcola confidenze. Specifica se durante il calcolo del punteggio vengono calcolate le confidenze.

Calcola punteggi di propensione grezza (solo obiettivi flag). Solo per i modelli con obiettivi di tipo flag, è possibile richiedere i punteggi di propensione grezza che indicano la verosimiglianza del risultato *vero* specificato per il campo obiettivo. Tali punteggi si aggiungono alla previsione standard e ai valori di confidenza. I punteggi di propensione regolata non sono disponibili. Per ulteriori informazioni, consultare l'argomento “Opzioni della scheda Analizza di un nodo Modelli” a pagina 34.

Accoda tutte le probabilità. Specifica se le probabilità di ogni categoria del campo di output vengono aggiunte a ogni record elaborato dal nodo. Se questa opzione non è selezionata, verrà aggiunta solo la probabilità della categoria prevista. Per un obiettivo nominale con tre categorie, per esempio, il risultato del calcolo del punteggio includerà una colonna per ciascuna delle tre categorie, oltre a una quarta colonna che indica la probabilità della categoria prevista. Per esempio, se le probabilità delle categorie *Rosso*, *Verde* e *Blu* sono rispettivamente 0.6, 0.3 e 0.1, la categoria prevista sarà *Rosso*, con una probabilità dello 0.6.

Calcola il punteggio convertendo in SQL nativo. Se questa opzione è selezionata, genera il codice SQL per calcolare in modo nativo il punteggio del modello all'interno dell'applicazione.

Nota: per i modelli multinomiali, la generazione di codice SQL non è disponibile se è stata selezionata l'opzione **Accoda tutte le probabilità** oppure —per i modelli con obiettivi nominali— se è stata selezionata l'opzione **Calcola confidenze**. La generazione SQL con i calcoli di confidenza è supportata solo per i modelli multinomiali con obiettivi di tipo flag. La generazione SQL non è disponibile per i modelli binomiali.

Modelli binomiali

Per i modelli binomiali, le confidenze e le probabilità sono sempre abilitate e le impostazioni che consentono di disabilitare tali opzioni non sono disponibili. La generazione SQL non è disponibile per i modelli binomiali. L'unica impostazione che è possibile modificare per i modelli binomiali è la possibilità di calcolare punteggi di propensione grezza. Come notato in precedenza per i modelli multinomiali, questo è valido solo per i modelli con obiettivi di tipo flag. Per ulteriori informazioni, consultare l'argomento “Opzioni della scheda Analizza di un nodo Modelli” a pagina 34.

Output avanzato del nugget del modello Logistica

L'output delle opzioni avanzate relative alla regressione logistica (anche nota come **regressione nominale**) contiene informazioni dettagliate sul modello stimato e sulle relative prestazioni. La maggior parte delle

informazioni contenute nell'output delle opzioni avanzate è piuttosto tecnica, quindi sono necessarie approfondite conoscenze tecniche dell'analisi di regressione logistica per interpretare correttamente tali dati.

Avvisi. Indica gli eventuali avvisi e i problemi potenziali relativi ai risultati.

Riepilogo dell'elaborazione dei casi. Elenca il numero di record elaborati, interrotti da ogni campo simbolico incluso nel modello.

Riepilogo fasi (facoltativo). Elenca gli effetti aggiunti o eliminati a ogni passaggio della creazione del modello quando si utilizza la selezione automatica dei campi.

Nota: visualizzata solo per i metodi Stepwise, In avanti, All'indietro o Stepwise All'indietro.

Cronologia iterazione (facoltativo). Visualizza la cronologia delle iterazioni delle stime dei parametri per ogni *n* iterazioni iniziando con la stima iniziale, dove *n* è il valore dell'intervallo di stampa. Per default vengono stampate tutte le iterazioni (n=1).

Informazioni sull'adattamento del modello (modelli multinomiali). Visualizza il test rapporto di verosimiglianza del modello (Finale) confrontandolo con un modello in cui tutti i coefficienti di parametro sono 0 (Solo intercettazione).

Classificazione (facoltativo). Visualizza la matrice dei valori dei campi di output effettivi e previsti con le percentuali.

Statistiche chi-quadrato sulla bontà di adattamento (facoltativo). Visualizza le statistiche del chi-quadrato sul rapporto di verosimiglianza e di Pearson. Queste statistiche verificano l'adattamento globale del modello ai dati di addestramento.

Bontà di adattamento Hosmer-Lemeshow (facoltativo). Mostra i risultati del raggruppamento dei casi in decili di rischio e del confronto della probabilità osservata con la probabilità attesa all'interno di ogni decile. Questa statistica della bontà di adattamento è più efficace della statistica della bontà di adattamento tradizionale utilizzata nei modelli multinomiali, soprattutto per i modelli con covariate continue e gli studi con campioni di dimensioni ridotte.

Pseudo R-quadrato (facoltativo). Visualizza le misure dell'adattamento del modello di Cox e Snell, Nagelkerke e l'*R*-quadrato di McFadden. Queste statistiche sono per alcuni aspetti analoghe alla statistica *R*-quadrato nella regressione lineare.

Misure di monotonicità (facoltativo). Mostra il numero di coppie concordanti, di coppie discordanti e di coppie pari merito nei dati, oltre alla percentuale del numero totale di coppie che ciascuna rappresenta. In questa tabella sono inoltre visualizzati D di Somers, Gamma di Goodman e Kruskal, tau-a di Kendall e l'indice di concordanza C.

Criteri di informazione (facoltativo). Mostra i criteri di informazioni di Akaike (AIC, Akaike's information criterion) e bayesiano di Schwarz (BIC, Bayesian information criterion).

Test rapporto di verosimiglianza (facoltativo). Mostra le statistiche che verificano se i coefficienti degli effetti del modello sono statisticamente diversi da 0. I campi di input significativi sono quelli con livelli di significatività molto bassi nell'output (etichettati *Sig.*).

Stime parametri (facoltativo). Visualizza le stime dei coefficienti di equazione, i relativi test, i rapporti odds derivati dagli stessi coefficienti contraddistinti dall'etichetta *Exp(B)* e gli intervalli di confidenza relativi ai rapporti odds.

Matrice di correlazione/covarianza asintotica (facoltativo). Visualizza le correlazioni e/o covarianze asintotiche delle stime dei coefficienti.

Frequenze previste e osservate (facoltativo). In ogni struttura di covariata, visualizza le frequenze previste e osservate relative al valore di ogni campo di output. Questa tabella può avere dimensioni anche molto grandi, soprattutto nei modelli con campi di input numerici. Se la tabella risultante ha dimensioni tali da comprometterne la corretta visualizzazione, viene omessa e viene visualizzato un avviso.

Nodo fattoriale/PCA

Il nodo fattoriale/PCA offre potenti tecniche di riduzione dei dati che consentono di diminuirne la complessità. Esistono due approcci simili ma distinti.

- **L'analisi dei componenti principali (PCA, Principal Components Analysis)** trova le combinazioni lineari dei campi di input che catturano meglio la varianza nell'intero insieme di campi, dove i componenti sono ortogonali (perpendicolari) l'uno rispetto all'altro. La PCA prende in esame qualsiasi varianza, sia quella condivisa sia quella univoca.
- L'**analisi fattoriale** tenta di identificare i concetti sottostanti, o **fattori**, che spiegano lo schema delle correlazioni all'interno dell'insieme di campi osservati. L'analisi fattoriale prende in esame solo la varianza condivisa. La varianza che interessa solo specifici campi non viene presa in considerazione nella stima del modello. Il nodo Fattoriale offre diversi metodi di analisi fattoriale.

Entrambi gli approcci mirano a trovare un numero ridotto di campi derivati che riassumono in modo efficace le informazioni presenti nell'insieme originale di campi.

Requisiti. In un modello fattoriale/PCA è possibile utilizzare solo campi numerici. Per stimare un'analisi fattoriale o PCA, è necessario che siano presenti uno o più campi con il ruolo impostato su *Input*. I campi con ruolo impostato su *Obiettivo*, *Entrambi* o *Nessuno* verranno ignorati, così come i campi non numerici.

Efficacia. L'analisi fattoriale e l'analisi PCA possono ridurre in modo efficiente la complessità dei dati senza pregiudicare troppo il contenuto delle informazioni. Queste tecniche possono semplificare la generazione di modelli più solidi e la cui esecuzione è più rapida di quanto non sarebbe possibile con i campi di input non elaborati.

Opzioni del modello di nodo fattoriale/PCA

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.

Utilizza dati partizionati. Se è definito un campo partizione, questa opzione garantisce che per la creazione del modello vengano utilizzati dati solo dalla partizione di addestramento.

Metodo di estrazione. Specificare il metodo da utilizzare per la riduzione dei dati.

- **Componenti principali.** Questo è il metodo di default che utilizza l'analisi PCA per trovare i componenti che riassumono i campi di input.
- **Minimi quadrati non ponderati.** Questo metodo di analisi fattoriale trova l'insieme di fattori che riproduce meglio lo schema di relazioni (correlazioni) tra i campi di input.
- **Minimi quadrati generalizzati.** Questo metodo di analisi fattoriale è simile ai minimi quadrati non ponderati, ma utilizza la ponderazione per de-enfatizzare i campi con molta varianza univoca (non condivisa).
- **Massima verosimiglianza.** Questo metodo di analisi fattoriale produce le equazioni fattoriali che più probabilmente hanno dato origine al modello di relazioni osservato (correlazioni) nei campi di input, in base ai presupposti relativi al formato di tali relazioni. In particolare, il metodo presume che i dati di addestramento seguano una distribuzione normale a più variabili.

- **Fattorizzazione asse principale.** Questo metodo di analisi fattoriale è molto simile al metodo di analisi dei componenti principali, tranne per il fatto che considera solo la varianza condivisa.
- **Fattorizzazione alfa.** Questo metodo di analisi fattoriale ritiene che i campi nell'analisi siano un campione dell'universo di potenziali campi di input. Massimizza l'affidabilità statistica dei fattori.
- **Fattorizzazione immagine.** Questo metodo di analisi fattoriale utilizza la stima dei dati per isolare la varianza comune e trovare i fattori che la descrivono.

Opzioni avanzate del nodo fattoriale/PCA

Le opzioni avanzate consentono agli utenti esperti di analisi fattoriale e PCA di ottimizzare il processo di addestramento. Per accedere alle opzioni avanzate, impostare la modalità su **Livello avanzato** nella scheda Livello avanzato.

Valori mancanti. Per default, IBM SPSS Modeler utilizza solo i record con valori validi per tutti i campi utilizzati nel modello (questo approccio viene a volte chiamato **eliminazione listwise** dei valori mancanti). Tuttavia, se la quantità di valori mancanti è notevole, è possibile che in questo modo venga eliminato un numero eccessivo di record e che pertanto i dati disponibili non siano sufficienti per generare un modello efficace. In tali casi, è possibile deselezionare l'opzione **Utilizza solo record completi**. IBM SPSS Modeler quindi prova ad utilizzare la maggior quantità di informazioni possibile per eseguire la stima del modello, inclusi i record in cui alcuni dei campi hanno valori mancanti. (questo approccio viene a volte chiamato **eliminazione pairwise** dei valori mancanti). Tuttavia, in alcune situazioni, l'utilizzo di record incompleti può portare a problemi di calcolo nella stima del modello.

Campi. Specificare se utilizzare la matrice di correlazione (impostazione di default) o la matrice di covarianza dei campi di input nella stima del modello.

Numero massimo di iterazioni per la convergenza. Specificare il numero massimo di iterazioni per la stima del modello.

Estrai fattori. Esistono due modi per selezionare il numero di fattori da estrarre dai campi di input.

- **Autovalori oltre.** Questa opzione manterrà tutti i fattori o i componenti con autovalori maggiori del criterio specificato. **Autovalori** misura la capacità di ciascun fattore o componente di riassumere la varianza nell'insieme di campi di input. Il modello manterrà tutti i fattori o componenti con autovalori maggiori del valore specificato quando si utilizza la matrice di correlazione. Quando si utilizza la matrice di covarianza, il criterio è il valore specificato moltiplicato per l'autovalore medio. Questa graduazione conferisce a questa opzione un significato simile per entrambi i tipi di matrice.
- **Numero massimo.** Questa opzione manterrà il numero di fattori o componenti specificato in ordine decrescente per gli autovalori. In altre parole, vengono mantenuti i fattori o i componenti che corrispondono a *n* autovalori massimi, dove *n* è il criterio specificato. Il criterio di estrazione di default è cinque fattori/componenti.

Formato matrice componente/fattore. Queste opzioni controllano il formato della matrice fattoriale (o matrice dei componenti per i modelli PCA).

- **Ordina valori.** Se questa opzione è selezionata, i caricamenti dei fattori nell'output del modello verranno ordinati numericamente.
- **Nascondi valori al di sotto di.** Se questa opzione è selezionata, i punteggi al di sotto della soglia specificata verranno nascosti nella matrice per semplificare la visualizzazione dello schema nella matrice.

Rotazione. Queste opzioni consentono di controllare il metodo di rotazione per il modello. Per ulteriori informazioni, consultare l'argomento “Opzioni di rotazione del nodo fattoriale/PCA” a pagina 182.

Opzioni di rotazione del nodo fattoriale/PCA

In molti casi, la rotazione matematica dell'insieme di fattori mantenuti può aumentare la loro utilità e, in particolare, può agevolarne l'interpretazione. Selezionare un metodo di rotazione:

- **Nessuna rotazione.** Questa è l'opzione di default. Non viene applicata nessuna rotazione.
- **Varimax.** Un metodo di rotazione ortogonale che riduce al minimo il numero di campi con elevati caricamenti su ciascun fattore. Semplifica l'interpretazione dei fattori.
- **Oblimin diretto.** Un metodo per la rotazione obliqua (non ortogonale). Quando **Delta** equivale a 0 (impostazione di default), le soluzioni sono oblique. Quando delta diventa negativo e aumenta in valore assoluto, i fattori cominciano a essere meno obliqui. Inserire un numero minore o uguale a 0,8 per sovrascrivere il valore di default.
- **Quartimax.** Un metodo ortogonale che riduce al minimo il numero di fattori necessari per spiegare ciascun campo. Semplifica l'interpretazione dei campi osservati.
- **Equamax.** Un metodo di rotazione che è una combinazione del metodo Varimax, che semplifica i fattori, e del metodo Quartimax, che semplifica i campi. Il numero di campi che pesano fortemente su un fattore e il numero di fattori necessari per spiegare un campo vengono ridotti al minimo.
- **Promax.** Una rotazione obliqua, che consente di correlare i fattori. Può essere calcolata più rapidamente di una rotazione Oblimin diretto, quindi può risultare utile nel caso di insiemi di dati di grandi dimensioni. **Kappa** controlla l'obliquità della soluzione, ovvero in che misura i fattori possono essere correlati.

Nugget del modello fattoriale/PCA

Un nugget del modello fattoriale/PCA rappresenta il modello di analisi e analisi dei componenti principali (PCA) creato da un nodo Fattoriale. Contengono tutte le informazioni intercettate dal modello addestrato, nonché le informazioni relative alle caratteristiche e alle prestazioni del modello.

Quando viene eseguito un flusso contenente un nodo Equazione fattoriale, tale nodo aggiunge un nuovo campo per ogni fattore o componente nel modello. I nomi dei nuovi campi derivano dal nome del modello, con l'aggiunta del prefisso *$F-* e del suffisso *-n*, dove *n* è il numero del fattore o del componente. Per esempio, se il modello è denominato *Fattore* e contiene tre fattori, i nuovi campi saranno denominati *$F-Fattore-1*, *$F-Fattore-2* e *$F-Fattore-3*.

Per comprendere meglio cosa è stato codificato dal modello fattoriale, è possibile eseguire qualche analisi in più a valle. Un modo utile per visualizzare il risultato del modello fattoriale consiste nel visualizzare le correlazioni tra fattori e campi di input utilizzando un nodo Statistiche. In questo modo vengono visualizzati quali campi di input hanno più peso su quali fattori e consentono di scoprire se i fattori hanno un'interpretazione o un significato sottostante.

È inoltre possibile valutare il modello fattoriale utilizzando le informazioni disponibili nell'output delle opzioni avanzate. Per visualizzare l'output delle opzioni avanzate, fare clic sulla scheda **Opzioni avanzate** del browser del nugget del modello. L'output delle opzioni avanzate contiene molte informazioni dettagliate ed è destinato agli utenti che possiedono un'approfondita conoscenza dell'analisi fattoriale o PCA. Per ulteriori informazioni, consultare l'argomento “Output avanzato del nugget del modello fattoriale/PCA” a pagina 183.

Equazioni del nugget del modello fattoriale/PCA

La scheda Modello di un nugget del modello fattoriale visualizza l'equazione del punteggio fattoriale per ogni fattore. I punteggi relativi a componenti e fattori vengono calcolati moltiplicando il valore di ogni campo di input in base al coefficiente e sommando i risultati.

Riepilogo del nugget del modello fattoriale/PCA

La scheda Riepilogo di un modello fattoriale visualizza il numero di fattori mantenuti nel modello Fattoriale, oltre ad informazioni aggiuntive sui campi e alle impostazioni utilizzate per generare il modello. Per ulteriori informazioni, consultare l'argomento “Esplorazione dei nugget del modello” a pagina 42.

Output avanzato del nugget del modello fattoriale/PCA

L'output della scheda Opzioni avanzate relativa all'analisi fattoriale contiene informazioni dettagliate sul modello stimato e sulle relative prestazioni. La maggior parte delle informazioni contenute nell'output delle opzioni avanzate è piuttosto tecnica, quindi sono necessarie approfondite conoscenze tecniche di analisi fattoriale per interpretare correttamente tali dati.

Avvisi. Indica gli eventuali avvisi e i problemi potenziali relativi ai risultati.

Comunalità. Mostra la proporzione della varianza di ogni campo spiegata dai fattori o dai componenti. *Iniziale* contiene le comunalità iniziali e l'insieme completo di fattori (il modello inizia con tanti fattori quanti sono i campi di input) e *Estrazione* contiene le comunalità basate sull'insieme di fattori mantenuti.

Varianza totale spiegata. Mostra la varianza totale spiegata dai fattori nel modello. *Autovalori iniziali* mostra la varianza spiegata dall'intero insieme di fattori iniziali. *Pesi dei fattori non ruotati* mostra la varianza spiegata dai fattori mantenuti nel modello. *Pesi dei fattori ruotati* mostra la varianza spiegata dai fattori ruotati. Si noti che nel caso delle rotazioni oblique, *Pesi dei fattori ruotati* mostra solo i fattori ruotati, non le percentuali della varianza.

Matrice dei fattori o dei componenti. Mostra le correlazioni tra campi di input e fattori non ruotati.

Matrice dei componenti o dei fattori ruotati. Mostra le correlazioni tra i campi di input e i fattori ruotati relativi alle rotazioni ortogonali.

Matrice dei modelli. Mostra le correlazioni parziali tra i campi di input e i fattori ruotati relativi alle rotazioni oblique.

Matrice della struttura. Mostra le correlazioni semplici tra i campi di input e i fattori ruotati relativi alle rotazioni oblique.

Matrice di correlazione dei fattori. Mostra correlazioni tra fattori relativi alle rotazioni oblique.

Nodo Discriminante

L'analisi discriminante crea un modello predittivo per l'appartenenza ai gruppi. Il modello è composto da una funzione discriminante (oppure, per più di due gruppi, da un insieme di funzioni discriminanti) basata su combinazioni lineari delle variabili predittore che forniscono la discriminazione ottimale tra i gruppi. Le funzioni vengono generate da un campione di casi di cui è nota l'appartenenza; le funzioni possono in seguito essere applicate ai nuovi casi che hanno misurazioni delle variabili predittore ma la cui appartenenza di gruppo è sconosciuta.

Esempio. Una società di telecomunicazioni può utilizzare l'analisi discriminante per classificare i clienti e suddividerli in gruppi in base ai dati di utilizzo. In questo modo, è possibile assegnare un punteggio ai potenziali clienti e concentrarsi su quelli che hanno maggiori probabilità di rientrare nei gruppi più significativi.

Requisiti. Sono necessari uno o più campi di input ed esattamente un campo obiettivo. L'obiettivo deve essere un campo categoriale (con un livello di misurazione *Flag* o *Nominale*) con tipo di archiviazione come stringa o numero intero. Se necessario, è possibile convertire l'archiviazione utilizzando un nodo

Riempimento o Ricava. I campi impostati su *Entrambe* o *Nessuna* verranno ignorati. È necessario che i tipi dei campi utilizzati nel modello siano completamente istanziati.

Efficacia. Sia l'analisi discriminante sia la regressione logistica sono modelli di classificazione adeguati. Tuttavia, l'analisi discriminante prevede un numero maggiore di presupposti relativi ai campi di input — ad esempio, che siano normalmente distribuiti e continui e che forniscano risultati migliori se tali requisiti sono soddisfatti, specialmente se il campione è di dimensioni ridotte.

Opzioni del modello di nodo Discriminante

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.

Utilizza dati partizionati. Se è definito un campo partizione, questa opzione garantisce che per la creazione del modello vengano utilizzati dati solo dalla partizione di addestramento.

Crea modelli di suddivisione. Crea un modello separato per ogni valore possibile dei campi di input che sono stati specificati come campi di suddivisione. Per ulteriori informazioni, consultare l'argomento "Creazione di modelli di suddivisione" a pagina 28.

Metodo. Per l'immissione dei predittori nel modello, sono disponibili le seguenti opzioni:

- **Invio.** Questo è il metodo di default che immette tutti i termini direttamente nell'equazione. I termini che non aumentano in modo significativo il potere predittivo del modello non vengono aggiunti.
- **Stepwise.** Il modello iniziale è il più semplice possibile e l'equazione non presenta termini di modello (ad eccezione della costante). A ogni fase vengono valutati i termini che ancora non sono stati aggiunti al modello e, se il migliore di essi aumenta notevolmente il potere predittivo del modello, viene aggiunto.

Nota: il metodo Stepwise ha una forte tendenza a sovradattare i dati di addestramento. Quando vengono utilizzati questi metodi, è particolarmente importante verificare la validità del modello risultante con un campione estratto per il test o con dati nuovi.

Opzioni avanzate del nodo Discriminante

Le opzioni avanzate consentono agli utenti esperti dell'analisi discriminante di ottimizzare il processo di addestramento. Per accedere alle opzioni avanzate, impostare **Modalità** su **Livello avanzato** nella scheda Livello avanzato.

Probabilità a priori. Questa opzione stabilisce se i coefficienti di classificazione debbano essere corretti per una conoscenza a priori dell'appartenenza ai gruppi.

- **Tutti i gruppi uguali.** Per tutti i gruppi vengono presunte uguali probabilità a priori; ciò non ha alcun effetto sui coefficienti.
- **Calcola dalle dimensioni dei gruppi.** Le dimensioni dei gruppi osservate nel campione determinano le probabilità a priori dell'appartenenza ai gruppi. Per esempio, se il 50% delle osservazioni incluse nell'analisi rientra nel primo gruppo, il 25% nel secondo e il 25% nel terzo, i coefficienti di classificazione vengono regolati in modo da aumentare la verosimiglianza dell'appartenenza al primo gruppo rispetto agli altri due.

Usa matrice di covarianza. È possibile decidere di classificare i casi utilizzando una matrice di covarianza entro i gruppi o una matrice di covarianza per gruppi separati.

- *Entro i gruppi*. La matrice di covarianza entro i gruppi viene utilizzato per classificare casi.
- *Gruppi separati*. Per classificare i casi vengono utilizzate le matrici di covarianza dei singoli gruppi. Dal momento che la classificazione è basata sulla funzione discriminante e non sui valori originali, questa opzione non è sempre equivalente alla discriminazione quadratica.

Output. Queste opzioni permettono di richiedere statistiche aggiuntive che verranno visualizzate nell'output avanzato del nugget del modello creato dal nodo. Per ulteriori informazioni, consultare l'argomento “Opzioni di output del nodo Discriminante”.

Controllo. Queste opzioni permettono di controllare i criteri per l'aggiunta e la rimozione di campi con il metodo di stima Stepwise. Se il metodo selezionato è Per blocchi, il pulsante è disattivato. Per ulteriori informazioni, consultare l'argomento “Opzioni di controllo del nodo Discriminante” a pagina 186.

Opzioni di output del nodo Discriminante

Selezionare l'output facoltativo che si desidera visualizzare nell'output avanzato del nugget del modello di regressione logistica. Per visualizzare l'output avanzato, individuare il nugget del modello e fare clic sulla scheda **Opzioni avanzate**. Per ulteriori informazioni, consultare l'argomento “Output avanzato del nugget del modello Discriminante” a pagina 187.

Descrittive. Le opzioni disponibili sono medie (incluse deviazioni standard), ANOVA univariate e test *M* di Box.

- *Medie*. Visualizza le medie totali e di gruppo, nonché le deviazioni standard per le variabili indipendenti.
- *ANOVA univariate*. Effettua l'analisi della varianza univariata a una via per verificare l'uguaglianza delle medie di gruppo per ciascuna variabile indipendente.
- *M di Box*. Un test per l'uguaglianza di matrici di covarianza di gruppo. Per dimensioni di campione sufficientemente elevate, un valore P non significativo vuol dire che non ci sono sufficienti prove che le matrici differiscano. Il test è sensibile a scostamenti dalla normalità multivariata.

Coefficienti funzioni. Le opzioni disponibili sono i coefficienti di classificazione di Fisher e i coefficienti non standardizzati.

- *Di Fisher*. Visualizza i coefficienti di Fisher della funzione di classificazione, che possono essere usati direttamente per la classificazione. Viene riprodotto un insieme separato di coefficienti di funzioni di classificazione per ciascun gruppo. Ogni caso viene assegnato al gruppo in cui ottiene il più alto punteggio discriminante (valore della funzione di classificazione).
- *Non standardizzati*. Visualizza i coefficienti della funzione discriminante non standardizzata.

Matrici. Le matrici dei coefficienti disponibili per le variabili indipendenti sono la matrice di correlazione entro i gruppi, la matrice di covarianza entro i gruppi, la matrice di covarianza per gruppi separati e la matrice di covarianza totale.

- *Correlazione entro i gruppi*. Visualizza la matrice di correlazione entro gruppi ottenuta mediando le matrici di covarianza di tutti i gruppi prima di calcolare le correlazioni.
- *Covarianza entro i gruppi*. Visualizza una matrice combinata di covarianza entro i gruppi, che potrebbe differire dalla matrice di covarianza totale. La matrice è ottenuta dalla media delle matrici di covarianza di tutti i gruppi.
- *Covarianza per gruppi separati*. Visualizza matrici di covarianza separate per ciascun gruppo.
- *Covarianza totale*. Visualizza una matrice di covarianza di tutti i casi come se provenissero da un unico campione.

Classificazione. Il seguente output è relativo ai risultati di classificazione.

- *Risultati per casi*. Visualizza per ciascun caso i codici del gruppo effettivo, del gruppo previsto, della probabilità a posteriori e del punteggio discriminante.
- *Tabella di riepilogo*. Il numero di casi assegnati in modo corretto e non corretto a ciascuno dei gruppi in base all'analisi discriminante. A volte detta "Matrice confusione".
- *Classificazione autoesclusiva*. Ogni caso viene classificato usando le funzioni ricavate da tutti i casi meno se stesso. Nota anche come classificazione "metodo U".

- *Mappa territoriale*. Un grafico dei confini usati per classificare i casi in gruppi in base ai valori di una funzione. I numeri corrispondono ai gruppi nei quali vengono classificati i casi. La media per ciascun gruppo è indicata da un asterisco all'interno dei suoi confini. La mappa non viene visualizzata se c'è una sola funzione discriminante.
- *Gruppi combinati*. Crea un grafico a dispersione per tutti i gruppi dei primi due valori di funzioni discriminanti. Se esiste una sola funzione, viene invece visualizzato un istogramma.
- *Gruppi separati*. Crea un grafico a dispersione per gruppi separati dei primi due valori di funzioni discriminanti. Se esiste una sola funzione, vengono invece visualizzati gli istogrammi.

Stepwise. Riepilogo dei passi visualizza le statistiche per tutte le variabili dopo ciascun passo; **F per distanze a coppie** visualizza una matrice di rapporti *F* a coppia per ciascuna coppia di gruppi. I rapporti *F* possono essere usati per eseguire test di significatività delle distanze Mahalanobis fra i gruppi.

Opzioni di controllo del nodo Discriminante

Metodo. Selezionare la statistica da utilizzare per l'inserimento o la rimozione di nuove variabili. Le alternative disponibili sono lambda di Wilks, varianza non spiegata, distanza di Mahalanobis, minimo rapporto *F* e *V* di Rao. Con il *V* di Rao è possibile specificare l'aumento minimo in *V* per la variabile da inserire.

- *Lambda di Wilks*. Un metodo di selezione delle variabili nell'analisi discriminante per passi che sceglie le variabili da inserire nell'equazione in base a quanto esse contribuiscono a minimizzare il Lambda di Wilks. Ad ogni passo viene inserita la variabile che minimizza il valore globale del Lambda di Wilks'.
- *Varianza non spiegata*. Ad ogni passo viene inserita la variabile che riduce al minimo la somma della variazione spiegata fra gruppi.
- *Distanza di Mahalanobis*. Una misura di quanto differiscano i valori di un caso per le variabili indipendenti, rispetto al valore medio di tutti i casi. Un'elevata distanza di Mahalanobis indica che un caso include valori estremi per una o più variabili indipendenti.
- *Rapporto F più piccolo*. Un metodo di selezione delle variabili nelle analisi per passi basato sulla massimizzazione di un rapporto F valutato tramite la distanza di Mahalanobis tra gruppi.
- *V di Rao*. Una misura delle differenze tra medie di gruppo. Detta anche traccia di Lawley-Hotelling. Ad ogni passo viene inserita la variabile che massimizza l'aumento della V di Rao. Dopo aver selezionato questa opzione, specificare l'incremento minimo che una variabile deve apportare per essere inserita nell'analisi.

Criteri. Le alternative disponibili sono **Usa valore di F** e **Usa probabilità di F**. Immettere i valori per l'immissione e la rimozione delle variabili.

- *Usa valore di F*. La variabile viene inserita nel modello se il relativo valore F è maggiore di quello di inserimento. La variabile viene altresì rimossa se il relativo valore F è minore di quello di rimozione. I valori di inserimento e di rimozione devono essere entrambi positivi e Inserimento deve essere maggiore di Rimozione Abbassando il valore di inserimento e/o alzando quello di rimozione Per rimuovere ulteriori variabili dal modello, aumentare il valore di rimozione.
- *Usa probabilità di F*. La variabile viene inserita nel modello se il livello di significatività del relativo valore di F è minore di quello di inserimento. La variabile viene altresì rimossa se il livello di significatività è maggiore di quello di rimozione. I valori di inserimento e di rimozione devono essere entrambi positivi e Inserimento deve essere minore di Rimozione. Alzando il valore di inserimento e/o abbassando quello di rimozione Per rimuovere ulteriori variabili dal modello, riduce il valore di rimozione.

Nugget del modello Discriminante

I nugget del modello Discriminante rappresentano le equazioni stimate dai nodi Discriminante. Contengono tutte le informazioni intercettate dal modello discriminante, nonché informazioni sulla performance e la struttura del modello.

Quando viene eseguito un flusso che contiene un nugget del modello Discriminante, il nodo aggiunge due nuovi campi contenenti la previsione del modello e la probabilità associata. I nomi dei nuovi campi derivano dal nome del campo di output di cui si sta eseguendo la previsione, a cui viene aggiunto il prefisso *$D-* per la categoria prevista e *$DP-* per la probabilità associata. Per esempio, per un campo di output denominato *colorepref*, i nuovi campi si chiameranno *$D-colorepref* e *$DP-colorepref*.

Generazione di un nodo Filtro. Il menu Genera consente di creare un nuovo nodo Filtro per passare campi di input basati sui risultati del modello.

Importanza predittore

Facoltativamente, nella scheda Modello, è possibile visualizzare anche un grafico che indica l'importanza relativa di ogni predittore nella stima del modello. In genere è opportuno concentrare la modellazione sui predittori più importanti e valutare se eliminare o ignorare quelli meno importanti. Notare che tale grafico è disponibile solo se è selezionata l'opzione **Calcola importanza predittore** nella scheda Analizza prima di generare il modello. Per ulteriori informazioni, consultare l'argomento "Importanza predittore" a pagina 43.

Output avanzato del nugget del modello Discriminante

L'output avanzato dell'analisi discriminante fornisce informazioni dettagliate sul modello stimato e sulla relativa performance. La maggior parte delle informazioni contenute nell'output avanzato è di natura piuttosto tecnica, quindi è necessaria una conoscenza approfondita dell'analisi discriminante per interpretare correttamente i dati. Per ulteriori informazioni, consultare l'argomento "Opzioni di output del nodo Discriminante" a pagina 185.

Impostazioni del nugget del modello Discriminante

La scheda Impostazioni del nugget del modello Discriminante consente di ottenere i punteggi di propensione quando si calcola il punteggio del modello. La scheda è disponibile solo per i modelli con obiettivi flag e solo dopo che il nugget del modello è stato aggiunto a un flusso.

Calcola punteggi di propensione grezza. Per i modelli con obiettivi di tipo flag (che restituiscono un risultato sì o nessuna previsione), è possibile richiedere i punteggi di propensione grezza che indicano la verosimiglianza del risultato vero specificato per il campo obiettivo. Tali punteggi si aggiungono ai valori di previsione e confidenza che potrebbero venire generati durante il calcolo del punteggio.

Calcola punteggi di propensione regolata. I punteggi di propensione grezza sono basati esclusivamente sui dati di addestramento e potrebbero essere eccessivamente ottimistici a causa della tendenza di molti modelli a sovradattare questi dati. Le propensioni regolate cercano di operare una compensazione esaminando le prestazioni del modello rispetto a una partizione di test o di convalida. Per utilizzare questa opzione è necessario che nel flusso sia definito un campo partizione e che nel nodo Modelli siano attivati i punteggi di propensione regolati, prima di generare il modello.

Scheda Riepilogo del nugget del modello Discriminante

La scheda Riepilogo di un nugget del modello Discriminante visualizza i campi e le impostazioni utilizzati per generare il modello. Inoltre, se è stato eseguito un nodo Analisi collegato a questo nodo Modelli, in questa sezione verranno visualizzate anche le informazioni sui risultati di tale analisi. Per informazioni generali relative all'utilizzo del browser del modello, consultare "Esplorazione dei nugget del modello" a pagina 42.

Nodo GenLin

Il modello lineare generalizzato amplia il modello lineare generale in modo che la variabile dipendente venga linearmente correlata ai fattori e alle covariate tramite una funzione di collegamento specifica. Inoltre, il modello consente alla variabile dipendente di avere una distribuzione non normale. Grazie alla formulazione estremamente generale del modello, copre i modelli statistici utilizzati di frequente, quali la regressione lineare per le risposte distribuite normalmente, i modelli logistici per i dati binari, i modelli

log-lineari per i dati dei conteggi, i modelli doppi logaritmici complementari per i dati di sopravvivenza censurati per intervallo, oltre a molti altri modelli statistici.

Esempi. Una compagnia di navigazione può utilizzare modelli lineari generalizzati per adattare una regressione di Poisson ai conteggi dei danni per diversi tipi di navi costruite in periodi differenti ed il modello risultante può consentire di determinare i tipi di navi più soggetti a subire danni.

Una compagnia di assicurazioni auto può utilizzare modelli lineari generalizzati per adattare una regressione gamma alle richieste di risarcimento danni per le auto ed il modello risultante può consentire di determinare i fattori che contribuiscono maggiormente all'ammontare della richiesta.

I ricercatori medici possono utilizzare modelli lineari generalizzati per adattare una regressione doppia logaritmica complementare a dati di sopravvivenza di censura per intervallo per prevedere la ricorrenza di una condizione medica.

I modelli lineari generalizzati funzionano mediante la creazione di un'equazione che collega i valori dei campi di input ai valori dei campi di output. Dopo essere stato generato, il modello può essere utilizzato per stimare i valori relativi a nuovi dati. Per ciascun record, viene calcolata la probabilità di appartenenza per ciascuna categoria di output possibile. La categoria obiettivo con la maggiore probabilità viene assegnata come valore di output previsto per quel record.

Requisiti. Sono necessari uno o più campi di input ed esattamente un campo obiettivo (che può avere livello di misurazione *Continuo* o *Flag*) con due o più categorie. È necessario che i tipi dei campi utilizzati nel modello siano completamente istanziati.

Efficacia. Il modello lineare generalizzato è estremamente flessibile, ma la scelta della struttura del modello non è un processo automatizzato e, pertanto, richiede un grado di familiarità con i propri dati che non è richiesto dagli algoritmi di tipo "black box".

Opzioni dei campi del nodo GenLin

Oltre alle opzioni personalizzate relative a obiettivi, input e partizioni generalmente disponibili nelle schede Campi del nodo di modellazione (consultare “Opzioni dei campi dei nodi Modelli” a pagina 31), il nodo GenLin offre le seguenti funzionalità aggiuntive.

Utilizza campo peso. Il parametro scala è un parametro del modello stimato correlato alla varianza della risposta. I pesi della scala sono valori "noti" che possono variare a seconda delle osservazioni. Se la variabile del peso della scala è stata specificata, il parametro scala, che è correlato alla varianza della risposta, viene diviso per la suddetta variabile per ciascuna osservazione. I record con valori del peso di scala minori o uguali a 0 oppure mancanti non vengono utilizzati nell'analisi.

Il campo obiettivo rappresenta il numero degli eventi che si verificano in un insieme di prove. Quando la risposta è una serie di eventi che si verifica in un insieme di prove, il campo obiettivo contiene il numero degli eventi ed è possibile selezionare un'ulteriore variabile contenente il numero delle prove. In alternativa, se il numero di prove è lo stesso per tutti i soggetti, è possibile specificare le prove utilizzando un valore fisso. Il numero di prove deve essere maggiore o uguale al numero di eventi di ciascun record. Gli eventi devono essere numeri interi non negativi, mentre le prove devono essere numeri interi positivi.

Opzioni del modello di nodo GenLin

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.

Utilizza dati partizionati. Se è definito un campo partizione, questa opzione garantisce che per la creazione del modello vengano utilizzati dati solo dalla partizione di addestramento.

Crea modelli di suddivisione. Crea un modello separato per ogni valore possibile dei campi di input che sono stati specificati come campi di suddivisione. Per ulteriori informazioni, consultare l'argomento “Creazione di modelli di suddivisione” a pagina 28.

Tipo di modello. Esistono due opzioni relative al tipo di modello da creare. **Solo effetti principali** fa sì che il modello includa solo i campi di input individualmente e non verifichi le interazioni (effetti di moltiplicazione) tra i vari campi di input. **Effetti principali e tutte le interazioni a due vie** include tutte le interazioni a due vie e gli effetti principali dei campi di input.

Offset. Il termine offset è un predittore "strutturale". Il suo coefficiente non è stimato dal modello ma si presume che abbia il valore 1; pertanto, i valori dell'offset vengono semplicemente aggiunti al predittore lineare dell'obiettivo. Ciò è particolarmente utile nei modelli di regressione di Poisson, nei quali ogni caso può avere diversi livelli di esposizione all'evento di interesse.

Per esempio, nella modellazione dei tassi di incidente per singolo conducente esiste una differenza significativa tra un autista responsabile di un incidente in tre anni di esperienza e un autista responsabile di un incidente in 25 anni. Il numero di incidenti può essere rappresentato come una risposta di Poisson o binomiale negativa con un collegamento log se il logaritmo naturale dell'esperienza del conducente viene incluso come termine di offset.

Altre combinazioni dei tipi di distribuzione e di collegamento richiederebbero altre trasformazioni della variabile di offset.

Nota: se si utilizza un campo offset variabile, il campo specificato non deve essere utilizzato anche come input. Se necessario, impostare il ruolo del campo offset su **Nessuno** in un nodo origine o Tipo a monte.

Categoria di base per obiettivo flag.

Per la risposta binaria, è possibile scegliere la categoria di riferimento della variabile dipendente. Ciò può influire su alcuni output, quali le stime dei parametri e i valori salvati, ma non dovrebbe modificare l'adattamento del modello. Per esempio, se la risposta binaria prende i valori 0 e 1:

- Per default, la procedura rende l'ultima categoria (con valore più alto), o 1, la categoria di riferimento. In questa situazione, le probabilità salvate del modello stimano la possibilità che un dato caso prenda il valore 0 e le stime dei parametri devono essere interpretate come relative alla verosimiglianza di categoria 0.
- Se si specifica la prima categoria (con valore più basso), o 0, come categoria di riferimento, le probabilità salvate del modello stimano la possibilità che un dato caso prenda il valore 1.
- Se si specifica la categoria personalizzata e per la variabile sono definite delle etichette, è possibile impostare la categoria di riferimento scegliendo un valore dall'elenco. Ciò può essere utile quando, durante la specifica di un modello, non si ricorda esattamente come era codificata una variabile particolare.

Includi l'intercettazione nel modello. L'intercettazione viene in genere inclusa nel modello. Se è possibile presumere che i dati passino attraverso l'origine, l'intercettazione può essere esclusa.

Opzioni avanzate del nodo GenLin

Le opzioni avanzate consentono agli utenti esperti dei modelli lineari generalizzati di ottimizzare il processo di addestramento. Per accedere alle opzioni avanzate, impostare **Modalità** su **Livello avanzato** nella scheda Livello avanzato.

Distribuzione e funzione di legame dei campi obiettivo

Distribuzione.

Questa selezione specifica la distribuzione della variabile dipendente. La possibilità di specificare una distribuzione non normale e una funzione di legame di non identità è il vantaggio essenziale del modello lineare generalizzato rispetto al modello lineare generale. Dal momento che è possibile combinare più distribuzioni e funzioni di collegamento e che molte di queste sono adatte a qualsiasi combinazione di dati, è generalmente consigliabile fare una valutazione teorica a priori oppure selezionare la combinazione che si ritiene possa essere più adatta.

- **Binomiale.** Questa distribuzione è adatta solo alle variabili che rappresentano una risposta o un numero di eventi binario.
- **Gamma.** Questa distribuzione è adatta alle variabili con valori di scala positivi asimmetrici verso valori positivi superiori. Se il valore dei dati è inferiore o uguale a 0 o è mancante, il caso corrispondente non viene usato nell'analisi.
- **Gaussiana inversa.** Questa distribuzione è adatta alle variabili con valori di scala positivi asimmetrici verso valori positivi superiori. Se il valore dei dati è inferiore o uguale a 0 o è mancante, il caso corrispondente non viene usato nell'analisi.
- **Binomiale negativa.** Questa distribuzione può essere vista come il numero di prove necessario per osservare *k* esiti positivi ed è adatta alle variabili con valori interi non negativi. Se il valore dei dati è un numero non intero, inferiore a 0 o mancante, il caso corrispondente non viene usato nell'analisi. Il valore fisso del parametro ausiliario della distribuzione binomiale negativa può essere qualsiasi numero maggiore o uguale a 0. Quando il parametro ausiliario è impostato su 0, l'utilizzo di questa distribuzione equivale all'utilizzo della distribuzione di Poisson.
- **Normale.** È adatta alle variabili di scala i cui valori si distribuiscono assumendo una forma simmetrica a campana intorno a un valore centrale (media). La variabile dipendente deve essere un valore numerico.
- **Poisson.** Questa distribuzione può essere considerata equivalente al numero di occorrenze di un evento desiderato in un intervallo di tempo fisso ed è indicata per le variabili con valori interi non negativi. Se il valore dei dati è un numero non intero, inferiore a 0 o mancante, il caso corrispondente non viene usato nell'analisi.
- **Tweedie.** Questa distribuzione è appropriata per le variabili che possono essere rappresentate da combinazioni di Poisson di distribuzioni gamma; la distribuzione risulta "mista" nel senso che combina proprietà di distribuzioni continue (accetta valori reali non negativi) e distribuzioni discrete (massa di probabilità positiva a un unico valore, 0). La variabile dipendente deve essere numerica, con valori di dati maggiori o uguali a zero. Se il valore di un dato è minore di zero o mancante, il caso corrispondente non viene utilizzato nell'analisi. Il valore fisso del parametro della distribuzione Tweedie può essere qualsiasi numero maggiore di uno e minore di due.
- **Multinomiale.** Questa distribuzione è appropriata per variabili che rappresentano una risposta ordinale. La variabile dipendente può essere un numero o una stringa e deve avere almeno due valori di dati validi distinti.

Funzioni di legame.

La funzione di legame è una trasformazione della variabile dipendente che consente la stima del modello. Sono disponibili le seguenti funzioni:

- **Identità.** $f(x)=x$. La variabile dipendente non viene trasformata. Questo legame può essere utilizzato con qualsiasi distribuzione.
- **Doppia logaritmica complementare.** $f(x)=\log(-\log(1-x))$. Questa funzione è indicata solo per la distribuzione binomiale.
- **Cauchit cumulativa.** $f(x) = \tan(\pi\ (x - 0.5))$, applicata alla probabilità cumulata di ciascuna categoria della risposta. È adatta solo con la distribuzione multinomiale.
- **Log-log complementare cumulativa.** $f(x)=\ln(-\ln(1-x))$, applicata alla probabilità cumulata di ciascuna categoria della risposta. È adatta solo con la distribuzione multinomiale.
- **Logit cumulativa.** $f(x)=\ln(x\ /\ (1-x))$, applicata alla probabilità cumulata di ciascuna categoria della risposta. È adatta solo con la distribuzione multinomiale.

- **Log-log negativa cumulativa.** $f(x)=-\ln(-\ln(x))$, applicata alla probabilità cumulata di ciascuna categoria della risposta. È adatta solo con la distribuzione multinomiale.
- **Probit cumulativa.** $f(x)=\Phi^{-1}(x)$, applicata alla probabilità cumulata di ciascuna categoria della risposta, dove Φ^{-1} è la funzione di distribuzione cumulativa normale standard inversa. È adatta solo con la distribuzione multinomiale.
- **Log.** $f(x)=\log(x)$. Questo legame può essere utilizzato con qualsiasi distribuzione.
- **Complemento log.** $f(x)=\log(1-x)$. Questa funzione è indicata solo per la distribuzione binomiale.
- **Logit.** $f(x)=\log(x\ /\ (1-x))$. Questa funzione è indicata solo per la distribuzione binomiale.
- **Binomiale negativa.** $f(x)=\log(x\ /\ (x+k^{-1}))$, dove k è il parametro ausiliario della distribuzione binomiale negativa. È adatta solo con la distribuzione binomiale negativa.
- **Doppia logaritmica negativa.** $f(x)=-\log(-\log(x))$. Questa funzione è indicata solo per la distribuzione binomiale.
- **Potenza odd.** $f(x)=[(x/(1-x))^{\alpha}-1]/\alpha$, se $\alpha \neq 0$. $f(x)=\log(x)$, se $\alpha=0$. α è la specifica numerica richiesta e deve essere un numero reale. Questa funzione è indicata solo per la distribuzione binomiale.
- **Probit.** $f(x)=\Phi^{-1}(x)$, dove Φ^{-1} è la funzione di distribuzione cumulativa normale standard inversa. Questa funzione è indicata solo per la distribuzione binomiale.
- **Potenza.** $f(x)=x^{\alpha}$, se $\alpha \neq 0$. $f(x)=\log(x)$, se $\alpha=0$. α è la specifica numerica richiesta e deve essere un numero reale. Questo legame può essere utilizzato con qualsiasi distribuzione.

Parametri. I controlli in questo gruppo consentono di specificare i valore dei parametri quando vengono scelte determinate opzioni di distribuzione.

- **Parametro per binomiale negativa.** Per la distribuzione binomiale negativa, scegliere se si desidera specificare un valore oppure consentire al sistema di fornire un valore stimato.
- **Parametro per Tweedie.** Per la distribuzione Tweedie, specificare un numero compreso tra 1.0 e 2.0 per il valore fisso.

  Stima dei parametri. I controlli in questo gruppo consentono di specificare i metodi di stima e di fornire valori iniziali per le stime dei parametri.

  – **Metodo.** È possibile selezionare un metodo di stima dei parametri. Scegliere tra Newton-Raphson, Fisher-scoring o un metodo ibrido in cui le iterazioni di Fisher-scoring vengono eseguite prima di passare al metodo di Newton-Raphson. Se durante la fase di Fisher-scoring del metodo ibrido, prima del raggiungimento del numero massimo di iterazioni di Fisher, viene raggiunta la convergenza, l'algoritmo continua con il metodo di Newton-Raphson.

  – **Metodo del parametro di scala.** È possibile selezionare il metodo di stima del parametro di scala. La massima verosimiglianza stima i parametri di scala utilizzando gli effetti del modello, benché questa opzione non sia valida se la risposta ha una distribuzione binomiale, di Poisson o binomiale negativa . Le opzioni di devianza e chi-quadrato di Pearson stimano il parametro di scala a partire dal valore di queste statistiche. In alternativa, è possibile specificare un valore fisso per il parametro di scala.
- **Matrice di covarianza.** Lo stimatore basato sul modello è il valore negativo dell'inverso generalizzato della matrice hessiana. Lo stimatore robusto (chiamato anche Huber/White/sandwich) è lo stimatore basato sul modello "corretto" che fornisce una stima uniforme della covarianza anche quando le specifiche della varianza e la funzione di collegamento sono errate.

Iterazioni. Queste opzioni permettono di controllare i parametri per la convergenza del modello. Per ulteriori informazioni, consultare l'argomento “Iterazioni dei modelli lineari generalizzati” a pagina 192.

Output. Queste opzioni permettono di richiedere statistiche aggiuntive che verranno visualizzate nell'output avanzato del nugget del modello creato dal nodo. Per ulteriori informazioni, consultare l'argomento “Output avanzato dei modelli lineari generalizzati” a pagina 192.

Tolleranza della singolarità. Le matrici singolari (o non invertibili) dispongono di colonne dipendenti in modo lineare, che possono causare gravi problemi all'algoritmo di stima. Anche le matrici quasi singolari

possono generare risultati imprecisi, pertanto la procedura tratta come singolare qualsiasi matrice il cui determinante sia inferiore alla tolleranza. Specificare un valore positivo.

Iterazioni dei modelli lineari generalizzati

È possibile impostare i parametri di convergenza per la stima del modello lineare generalizzato.

Iterazioni. Sono disponibili le seguenti opzioni:

- **Numero massimo di iterazioni.** Numero massimo di iterazioni eseguite dall'algoritmo. Specifica un intero non negativo.
- **Massimo numero di dimezzamenti.** Ad ogni iterazione, la dimensione di passo viene ridotta di un fattore di 0,5 finché la log-verosimiglianza non aumenta o non si raggiunge il numero massimo di dimezzamenti. Specificare un intero positivo.
- **Verificare la separazione dei punti dati.** Quando è selezionato, l'algoritmo esegue dei test per garantire che le stime dei parametri abbiano valori univoci. La separazione si verifica quando la procedura può generare un modello che classifica correttamente ogni caso. Questa opzione è disponibile per risposte binomiali con formato binario .

Criteri di convergenza. Sono disponibili le seguenti opzioni

- **Convergenza parametri.** Quando è selezionato, l'algoritmo si arresta dopo un'iterazione nella quale la variazione assoluta o relativa nelle stime dei parametri è minore del valore specificato, che deve essere positivo.
- **Convergenza verosimiglianza logaritmica.** Quando è selezionato, l'algoritmo si arresta dopo un'iterazione nella quale la variazione assoluta o relativa nella funzione di log-verosimiglianza è minore del valore specificato, che deve essere positivo.
- **Convergenza hessiana.** Per la specifica assoluta, si presume la convergenza se una statistica basata sulla convergenza hessiana è minore del valore positivo specificato. Per la specifica relativa, si presume la convergenza se la statistica è minore del prodotto tra il valore positivo specificato e il valore assoluto della log-verosimiglianza.

Output avanzato dei modelli lineari generalizzati

Selezionare l'output facoltativo che si desidera visualizzare nell'output avanzato del nugget del modello lineare generalizzato. Per visualizzare l'output avanzato, individuare il nugget del modello e fare clic sulla scheda **Opzioni avanzate**. Per ulteriori informazioni, consultare l'argomento “Output avanzato del nugget del modello GenLin” a pagina 194.

È disponibile il seguente output:

- **Riepilogo dell'elaborazione dei casi.** Visualizza il numero e la percentuale dei casi inclusi ed esclusi dall'analisi e la tabella di riepilogo dei dati correlati.
- **Statistiche descrittive.** Visualizza le statistiche descrittive e le informazioni di riepilogo su variabile dipendente, covariate e fattori.
- **Informazioni sul modello.** Visualizza il nome dell'insieme di dati, la variabile dipendente o le variabili eventi e prove, la variabile offset, la variabile peso di scala, la distribuzione della probabilità e la funzione di legame.
- **Statistiche della bontà di adattamento.** Visualizza la devianza e la devianza scalata, il chi-quadrato di Pearson e il chi-quadrato di Pearson scalato, la log-verosimiglianza, il criterio di informazione di Akaike (AIC), l'AIC corretto in campioni finiti (AICC), il criterio di informazione bayesiano (BIC) e l'AIC coerente (CAIC).
- **Statistiche di riepilogo del modello.** Visualizza i test di adattamento del modello, incluse le statistiche del rapporto di verosimiglianza per il test omnibus di adattamento del modello e le statistiche per i contrasti di tipo I o III di ogni effetto.

- **Stime dei parametri.** Visualizza le stime dei parametri e le statistiche di test e gli intervalli di confidenza corrispondenti. Se lo si desidera, è possibile visualizzare le stime dei parametri esponenziati oltre alle stime dei parametri grezzi.
- **Matrice di covarianza per le stime dei parametri.** Visualizza la matrice di covarianza dei parametri stimati.
- **Matrice di correlazione per le stime dei parametri.** Visualizza la matrice di correlazione dei parametri stimati.
- **Matrici del coefficiente di contrasto (L).** Visualizza i coefficienti di contrasto degli effetti di default e delle medie marginali stimate, se richiesto nella scheda Medie marginali.
- **Funzioni stimabili generali.** Visualizza le matrici per la generazione delle matrici del coefficiente di contrasto (L).
- **Cronologia iterazioni.** Visualizza la cronologia delle iterazioni per le stime dei parametri e la log-verosimiglianza e stampa l'ultima valutazione del vettore gradiente e della matrice hessiana. La tabella della cronologia delle iterazioni visualizza le stime dei parametri per per ciascuna iterazione *n* esima che inizia con l'iterazione 0esima (stime iniziali), dove *n* è il valore dell'intervallo di stampa. Se è richiesta la cronologia delle iterazioni, l'ultima iterazione viene sempre visualizzata indipendentemente da *n*.
- **Test del moltiplicatore di Lagrange.** Visualizza le statistiche del test del moltiplicatore di Lagrange per valutare la validità di un parametro di scala calcolato utilizzando la devianza, o chi-quadrato di Pearson, o impostato su un numero fisso per le distribuzioni normale, gamma e gaussiana inversa. Per la distribuzione binomiale negativa, viene verificato il parametro ausiliario fisso.

Effetti del modello. Sono disponibili le seguenti opzioni:

- **Tipo di analisi.** Specificare il tipo di analisi da eseguire. L'analisi di tipo I è generalmente appropriata quando si hanno dei motivi a priori per ordinare i predittori nel modello, mentre l'analisi di tipo III trova un'applicazione più generale. Le statistiche di Wald o rapporti di verosimiglianza sono calcolati in base alla selezione nel gruppo di statistiche chi-quadrato.
- **Intervalli di confidenza.** Specificare un livello di confidenza maggiore di 50 e minore di 100. Gli intervalli di Wald si basano sulla presunzione che i parametri abbiano una distribuzione normale asintotica. Gli intervalli di verosimiglianza dei profili sono più precisi ma possono essere impegnativi in termini di calcoli. Il livello di tolleranza per gli intervalli di verosimiglianza dei profili è il criterio usato per arrestare l'algoritmo iterativo utilizzato per calcolare gli intervalli.
- **Funzione di verosimiglianza logaritmica.** Controlla il formato di visualizzazione della funzione di log-verosimiglianza. La funzione completa include un termine aggiuntivo costante rispetto alla stima dei parametri; non ha effetti sulla stima dei parametri ed è escluso dalla visualizzazione in alcuni prodotti software.

Nugget del modello GenLin

Un nugget del modello GenLin rappresenta le equazioni stimate da un nodo GenLin. Esso contiene tutte le informazioni intercettate dal modello, nonché informazioni sulla performance e la struttura del modello.

Quando viene eseguito un flusso contenente un nugget del modello GenLin, il nodo aggiunge nuovi campi il cui contenuto dipende dalla natura del campo obiettivo:

- **Target indicatore.** Aggiunge campi contenenti la categoria prevista e la probabilità associata, nonché le probabilità per ciascuna categoria. I nomi dei primi due nuovi campi derivano dal nome del campo di output di cui si sta eseguendo la previsione, a cui viene aggiunto il prefisso *$G-* per la categoria prevista e *$GP-* per la probabilità associata. Per esempio, per un campo di output denominato *default*, i nuovi campi si chiameranno *$G-default* e *$GP-default*. Gli ultimi due campi aggiuntivi vengono denominati in base ai valori del campo di output, con l'aggiunta del prefisso *$GP-*. Per esempio, se i valori validi di *default* sono *Sì* e *No*, i nuovi campi si chiameranno rispettivamente *$GP-Sì* e *$GP-No*.
- **Target continuo.** Aggiunge campi contenenti la media prevista e l'errore standard.

- **Target continuo, che rappresenta il numero di eventi in una serie di prove.** Aggiunge campi contenenti la media prevista e l'errore standard.
- **Target ordinale.** Aggiunge i campi che contengono la categoria prevista e la probabilità associata per ciascun valore dell'insieme ordinato. I nomi dei campi derivano dal valore dell'insieme ordinato di cui si esegue la previsione, con i prefissi *$G-* per la categoria prevista e *$GP-* per la probabilità associata.

Generazione di un nodo Filtro. Il menu Genera consente di creare un nuovo nodo Filtro per passare campi di input basati sui risultati del modello.

Importanza predittore

Facoltativamente, nella scheda Modello, è possibile visualizzare anche un grafico che indica l'importanza relativa di ogni predittore nella stima del modello. In genere è opportuno concentrare la modellazione sui predittori più importanti e valutare se eliminare o ignorare quelli meno importanti. Notare che tale grafico è disponibile solo se è selezionata l'opzione **Calcola importanza predittore** nella scheda Analizza prima di generare il modello. Per ulteriori informazioni, consultare l'argomento "Importanza predittore" a pagina 43.

Output avanzato del nugget del modello GenLin

L'output avanzato dei modelli lineari generalizzati contiene informazioni dettagliate sul modello stimato e sulla relativa performance. La maggior parte delle informazioni contenute nell'output delle opzioni avanzate è piuttosto tecnica, quindi sono necessarie approfondite conoscenze tecniche di questo tipo di analisi per interpretare correttamente tali dati. Per ulteriori informazioni, consultare l'argomento "Output avanzato dei modelli lineari generalizzati" a pagina 192.

Impostazioni del nugget del modello GenLin

La scheda Impostazioni del nugget del modello GenLin consente di ottenere i punteggi di propensione quando si calcola il punteggio del modello. La scheda è disponibile solo per i modelli con obiettivi flag e solo dopo che il nugget del modello è stato aggiunto a un flusso.

Calcola punteggi di propensione grezza. Per i modelli con obiettivi di tipo flag (che restituiscono un risultato sì o nessuna previsione), è possibile richiedere i punteggi di propensione grezza che indicano la verosimiglianza del risultato vero specificato per il campo obiettivo. Tali punteggi si aggiungono ai valori di previsione e confidenza che potrebbero venire generati durante il calcolo del punteggio.

Calcola punteggi di propensione regolata. I punteggi di propensione grezza sono basati esclusivamente sui dati di addestramento e potrebbero essere eccessivamente ottimistici a causa della tendenza di molti modelli a sovradattare questi dati. Le propensioni regolate cercano di operare una compensazione esaminando le prestazioni del modello rispetto a una partizione di test o di convalida. Per utilizzare questa opzione è necessario che nel flusso sia definito un campo partizione e che nel nodo Modelli siano attivati i punteggi di propensione regolati, prima di generare il modello.

Scheda Riepilogo di un nugget del modello GenLin

La scheda Riepilogo di un nugget del modello GenLin visualizza i campi e le impostazioni utilizzati per generare il modello. Inoltre, se è stato eseguito un nodo Analisi collegato a questo nodo Modelli, in questa sezione verranno visualizzate anche le informazioni sui risultati di tale analisi. Per informazioni generali relative all'utilizzo del browser del modello, consultare "Esplorazione dei nugget del modello" a pagina 42.

Modelli misti lineari generalizzati

Nodo GLMM

Utilizzare questo nodo per creare un modello misto lineare generalizzato (GLMM).

Modelli misti lineari generalizzati

I modelli misti lineari generalizzati estendono il modello lineare in modo che:

- L'obiettivo venga linearmente correlato ai fattori e alle covariate tramite una funzione di collegamento specifica.
- L'obiettivo possa avere una distribuzione non normale.
- Le osservazioni possano essere correlate.

I modelli misti lineari generalizzati includono un'ampia gamma di modelli, dalla regressione lineare semplice ai modelli multilivello complessi per i dati longitudinali non normali.

Esempi. Un provveditorato agli studi regionale può utilizzare il modello misto lineare generalizzato per stabilire se un metodo di insegnamento sperimentale è efficace nel migliorare i risultati in matematica degli studenti. Gli studenti della stessa classe devono essere correlati, in quanto hanno lo stesso insegnante, e anche le classi nella stessa scuola possono essere correlate, per cui è possibile includere effetti casuali al livello della classe e della scuola per tenere conto di diverse fonti di variabilità. Per ulteriori informazioni, consultare l'argomento .

I ricercatori medici possono utilizzare un modello misto lineare generalizzato per stabilire se un nuovo farmaco anticonvulsivo può ridurre la frequenza delle crisi epilettiche in un paziente. Le misurazioni ripetute sullo stesso paziente, tipicamente, sono positivamente correlate, pertanto un modello misto con alcuni effetti casuali dovrebbe essere appropriato. Il campo obiettivo, il numero di crisi, accetta valori interi positivi, per cui un modello misto lineare generalizzato con una distribuzione di Poisson e collegamento log potrebbe essere appropriato. Per ulteriori informazioni, consultare l'argomento .

I dirigenti di una società fornitrice di servizi TV, telefono e Internet via cavo possono servirsi di un modello misto lineare generalizzato per scoprire più informazioni in merito ai potenziali clienti. Poiché le risposte possibili hanno livelli di misurazione nominale, gli analisti della società utilizzano un modello misto logit generalizzato con intercettazione causale per acquisire la correlazione tra le risposte alle domande relative all'utilizzo dei vari tipi di servizio (tv, telefono, internet) nelle risposte di un determinato partecipante all'indagine. Per ulteriori informazioni, consultare l'argomento .

La scheda Struttura dati consente di specificare le relazioni strutturali tra i record di un insieme di dati quando le osservazioni sono correlate. Se i record dell'insieme di dati rappresentano osservazioni indipendenti, non è necessario specificare alcun valore nella scheda.

Soggetti. La combinazione di valori dei campi categoriali specificati deve definire in modo univoco i soggetti all'interno dell'insieme di dati. Per esempio, un singolo campo *ID paziente* dovrebbe essere sufficiente a definire i soggetti in un ospedale, ma la combinazione di *ID ospedale* e *ID paziente* potrebbe essere necessaria qualora i numeri di identificazione dei pazienti non fossero univoci nei diversi ospedali. Se sono previste più misurazioni, vengono registrate più osservazioni per ciascun soggetto, quindi è possibile che ciascun soggetto occupi più record nell'ambito dell'insieme dei dati.

Un **soggetto** è un'unità di osservazione che può essere considerata indipendente rispetto ad altri soggetti. Ad esempio, i risultati della misurazione della pressione sanguigna di un paziente in uno studio medico sono da considerarsi indipendenti rispetto alle misurazioni effettuate sugli altri pazienti. La definizione dei soggetti è particolarmente importante nel caso in cui esistano misurazioni ripetute per ciascun soggetto e si desideri modellare la correlazione tra queste osservazioni. Ci si potrebbe aspettare, per esempio, che le letture della pressione sanguigna di un paziente effettuate in occasione di visite al medico consecutive siano correlate.

Tutti i campi specificati come Soggetti nella scheda Struttura dati vengono utilizzati per definire i soggetti per la struttura di covarianza residua e fornire l'elenco dei campi possibili per la definizione dei soggetti per le strutture di covarianza ad effetti casuali nel Blocco effetti casuali.

Misure ripetute. I campi specificati qui vengono utilizzati per identificare le osservazioni ripetute. Per esempio, una singola variabile *Settimana* potrebbe identificare 10 settimane di osservazioni in uno studio medico, mentre *Mese* e *Giorno* potrebbero essere utilizzati insieme per identificare le osservazioni giornaliere nel corso di un anno.

Definisci gruppi covarianza per. I campi categorici specificati in questo punto definiscono insiemi indipendenti di parametri di covarianza a effetti ripetuti; uno per ciascuna categoria definita dalla classificazione incrociata dei campi di raggruppamento. Tutti i soggetti presentano lo stesso tipo di covarianza; i soggetti all'interno dello stesso gruppo di covarianza presentano gli stessi valori per i parametri.

Tipo di covarianza ripetuta. Specifica la struttura di covarianza per i residui. Sono disponibili le seguenti strutture:

- Autoregressivo di primo ordine (AR1)
- Autoregressivo a media mobile (1,1) (ARMA11)
- Simmetria composta
- Diagonale
- Identità scalata
- Toeplitz
- Non strutturato
- Componenti della varianza

Obiettivo: Queste impostazioni definiscono l'obiettivo, la sua distribuzione e la sua relazione con i predittori attraverso la funzione di collegamento.

Obiettivo. L'obiettivo è obbligatorio. Può avere qualsiasi livello di misurazione e il livello di misurazione dell'obiettivo vincola le distribuzioni e le funzioni di collegamento appropriate.

- **Usa numero di prove come denominatore.** Se la risposta obiettivo rappresenta il numero di eventi di un insieme di prove, il campo obiettivo contiene il numero di eventi ed è possibile selezionare un ulteriore campo che contenga il numero di prove. Per esempio, quando si testa un nuovo pesticida è possibile esporre dei campioni di formiche a diverse concentrazioni di pesticida e quindi registrare il numero di formiche uccise e il numero di formiche esposte in ogni campione. In questo caso, il campo che registra il numero di formiche uccise deve essere specificato come campo obiettivo (eventi) e il campo che registra il numero di formiche in ogni campione deve essere specificato come campo prove. Se il numero di formiche è lo stesso per ogni campione, il numero di prove deve essere specificato utilizzando un valore fisso.

  Il numero di prove deve essere maggiore o uguale al numero di eventi di ciascun record. Gli eventi devono essere numeri interi non negativi, mentre le prove devono essere numeri interi positivi.
- **Personalizza categoria di riferimento.** Per un obiettivo categoriale, è possibile scegliere la categoria di riferimento. Questa operazione può influire su un determinato output, ad esempio quello delle stime dei parametri, ma non dovrebbe modificare l'adattamento del modello. Per esempio, se l'obiettivo assume i valori 0, 1 e 2, per impostazione predefinita, la procedura imposta l'ultima categoria (dal valore più alto), o 2, come categoria di riferimento. In questa situazione, le stime dei parametri devono essere interpretate come relative alla verosimiglianza della categoria 0 o 1 *relativa* alla verosimiglianza della categoria 2. Se si specifica una categoria personalizzata e l'obiettivo ha delle etichette definite, è possibile impostare la categoria di riferimento scegliendo un valore dall'elenco. Questo può risultare comodo quando, durante un'operazione di specifica di un modello, non ci si ricorda esattamente del modo in cui è stato codificato un determinato campo.

Relazione e distribuzione obiettivo (Collegamento) con modello lineare. Dati i valori dei predittori, il modello prevede che la distribuzione dei valori dell'obiettivo segua la forma specificata e che i valori obiettivo siano linearmente correlati ai predittori attraverso la funzione di collegamento specificata. Sono disponibili dei collegamenti per diversi modelli comuni oppure è possibile scegliere un'impostazione

Personalizzata se esiste una particolare combinazione delle funzioni di distribuzione e collegamento che si desidera adattare e che non è presente nell'elenco breve.

- **Modello lineare.** Specifica una distribuzione normale con un collegamento identità, utile quando l'obiettivo può essere previsto con un modello di regressione lineare o ANOVA.
- **Regressione gamma.** Specifica una distribuzione gamma con collegamento log da utilizzare quando l'obiettivo contiene tutti valori positivi e propende verso valori maggiori.
- **Loglineare.** Specifica una distribuzione di Poisson con collegamento log da utilizzare quando l'obiettivo rappresenta un conteggio di occorrenze in un determinato periodo di tempo.
- **Regressione binomiale negativa.** Specifica una distribuzione binomiale negativa con collegamento log da utilizzare quando l'obiettivo e il denominatore rappresentano il numero di prove necessarie per osservare i successi *k*.
- **Regressione logistica multinomiale.** Specifica una distribuzione multinomiale da utilizzare quando l'obiettivo è una risposta multi-categoria. Utilizza un collegamento logit cumulativo (risultati ordinali) o un collegamento logit generalizzato (risposte nominali multi-categoria).
- **Regressione logistica binaria.** Specifica una distribuzione binomiale con collegamento logit da utilizzare quando l'obiettivo è una risposta binaria prevista da un modello di regressione logistica.
- **Probit binaria.** Specifica una distribuzione binomiale con collegamento probit da utilizzare quando l'obiettivo è una risposta binaria con una distribuzione normale sottostante.
- **Sopravvivenza di censura per intervallo.** Specifica una distribuzione binomiale con collegamento log-log complementare, utile nell'analisi di sopravvivenza quando alcune osservazioni non includono un evento di terminazione.

Distribuzione

Questa selezione specifica la distribuzione dell'obiettivo. La possibilità di specificare una distribuzione non normale e una funzione di collegamento senza identità è uno dei principali vantaggi offerti dal modello misto lineare generalizzato rispetto a quello misto lineare. Dal momento che è possibile combinare più distribuzioni e funzioni di collegamento e che molte di queste sono adatte a qualsiasi combinazione di dati, è generalmente consigliabile fare una valutazione teorica a priori oppure selezionare la combinazione che si ritiene possa essere più adatta.

- **Binomiale.** Questa distribuzione è indicata solo per un obiettivo che rappresenta una risposta binaria o un numero di eventi.
- **Gamma.** Questa distribuzione è indicata per un obiettivo con valori di scala positivi che propendono verso valori positivi maggiori. Se il valore dei dati è inferiore o uguale a 0 o è mancante, il caso corrispondente non viene usato nell'analisi.
- **Gaussiana inversa.** Questa distribuzione è indicata per un obiettivo con valori di scala positivi che propendono verso valori positivi maggiori. Se il valore dei dati è inferiore o uguale a 0 o è mancante, il caso corrispondente non viene usato nell'analisi.
- **Multinomiale.** Questa distribuzione è appropriata per un obiettivo che rappresenta una risposta a più categorie. La forma del modello varia a seconda del livello di misurazione dell'obiettivo.

  Un obiettivo **nominale** produrrà un modello multinomiale nominale in cui per ogni categoria dell'obiettivo (ad eccezione della categoria di riferimento) viene effettuata la stima di un insieme separato di parametri del modello. Le stime dei parametri per un dato predittore mostrano la relazione tra quel predittore e la probabilità di ogni categoria dell'obiettivo, relativamente alla categoria di riferimento.

  Un obiettivo **ordinale** produrrà un modello multinomiale ordinale in cui il termine di intercettazione tradizionale viene sostituito con un insieme di parametri di **soglia** correlati alla probabilità cumulativa delle categorie obiettivo.
- **Binomiale negativa.** La regressione binomiale negativa utilizza una distribuzione binomiale con collegamento log da utilizzare quando l'obiettivo rappresenta un conteggio di occorrenze con varianza elevata.

- **Normale.** Questa distribuzione è indicata per un target continuo i cui valori presentano una distribuzione simmetrica a forma di campana intorno al valore centrale (medio).
- **Poisson.** Questa distribuzione può essere considerata equivalente al numero di occorrenze di un evento desiderato in un intervallo di tempo fisso ed è indicata per le variabili con valori interi non negativi. Se il valore dei dati è un numero non intero, inferiore a 0 o mancante, il caso corrispondente non viene usato nell'analisi.

Funzioni di collegamento

La funzione di collegamento è la trasformazione dell'obiettivo che permette di stimare il modello. Sono disponibili le seguenti funzioni:

- **Identità.** $f(x)=x$. L'obiettivo non viene trasformato. Questa funzione di collegamento può essere usata per tutti i tipi di distribuzioni, tranne che per quella multinomiale.
- **Doppia logaritmica complementare.** $f(x)=\log(-\log(1-x))$. È adatta solo con la distribuzione binomiale o multinomiale.
- **Cauchit.** $f(x) = \tan(\pi\ (x - 0.5))$. È adatta solo con la distribuzione binomiale o multinomiale.
- **Log.** $f(x)=\log(x)$. Questa funzione di collegamento può essere usata per tutti i tipi di distribuzioni, tranne che per quella multinomiale.
- **Complemento log.** $f(x)=\log(1-x)$. Questa funzione è indicata solo per la distribuzione binomiale.
- **Logit.** $f(x)=\log(x\ /\ (1-x))$. È adatta solo con la distribuzione binomiale o multinomiale.
- **Doppia logaritmica negativa.** $f(x)=-\log(-\log(x))$. È adatta solo con la distribuzione binomiale o multinomiale.
- **Probit.** $f(x)=\Phi^{-1}(x)$, dove Φ^{-1} è la funzione di distribuzione cumulativa normale standard inversa. È adatta solo con la distribuzione binomiale o multinomiale.
- **Potenza.** $f(x)=x^{\alpha}$, se $\alpha \neq 0$. $f(x)=\log(x)$, se $\alpha=0$. α è la specifica numerica richiesta e deve essere un numero reale. Questa funzione di collegamento può essere usata per tutti i tipi di distribuzioni, tranne che per quella multinomiale.

Effetti fissi: I fattori a effetti fissi vengono generalmente considerati come campi i cui valori di interesse sono integralmente rappresentati nell'insieme di dati e possono essere utilizzati per il calcolo del punteggio. Per impostazione predefinita, i campi con il ruolo di input predefinito che non sono specificati in altri punti della finestra di dialogo vengono immessi nella sezione degli effetti fissi del modello. I campi categoriali (flag, nominali e ordinali) vengono utilizzati come fattori nel modello e i campi continui vengono utilizzati come covariate.

Immettere gli effetti nel modello selezionando uno o più campi nell'elenco di sorgenti e trascinandoli nell'elenco degli effetti. Il tipo di effetto creato dipende dall'area sensibile nella quale si rilascia la selezione.

- **Principale.** I campi rilasciati vengono visualizzati come effetti principali separati in fondo all'elenco degli effetti.
- **2 vie.** Tutte le possibili coppie dei campi rilasciati vengono visualizzate come interazioni a 2 vie in fondo all'elenco degli effetti.
- **3 vie.** Tutti i possibili gruppi di tre dei campi rilasciati vengono visualizzati come interazioni a 3 vie in fondo all'elenco degli effetti.
- ***.** La combinazione di tutti i campi rilasciati viene visualizzata come una singola interazione nella parte inferiore dell'elenco degli effetti.

I pulsanti a destra del generatore di effetti consentono di eseguire diverse azioni.

Tabella 10. Descrizioni dei pulsanti del generatore di effetti.

Icona	Descrizione
	Eliminare i termini del modello a effetti fissi selezionando i termini che si desidera eliminare e facendo clic sul pulsante di eliminazione.
	Riordinare i termini all'interno del modello a effetti fissi selezionando i termini che si desidera riordinare e facendo clic sulla freccia rivolta verso l'alto o verso il basso.
	Aggiungere termini nidificati al modello utilizzando la finestra di dialogo "Aggiungi un termine personalizzato", facendo clic sul pulsante Aggiungi un termine personalizzato.

Includi intercettazione. Generalmente, l'intercettazione è inclusa nel modello. Se è possibile presumere che i dati passino attraverso l'origine, l'intercettazione può essere esclusa.

Aggiungi un termine personalizzato: Questa procedura consente di costruire termini nidificati per il modello. I termini nidificati sono utili per modellare l'effetto di un fattore o di una covariata i cui valori non interagiscono con i livelli di un altro fattore. Per esempio, una catena di supermercati può seguire le abitudini di spesa dei propri clienti in più negozi. Poiché ogni cliente frequenta un solo negozio, l'effetto *Cliente* può definirsi **nidificato** all'interno dell'effetto *Negozio*.

È inoltre possibile includere effetti di interazione, ad esempio termini polinomiali che interessano la stessa covariata, o aggiungere più livelli di nidificazione al termine nidificato.

Limitazione. I termini nidificati sono sottoposti alle seguenti restrizioni:

- Tutti i fattori compresi in un'interazione devono essere univoci. Di conseguenza, se *A* è un fattore, non è possibile specificare *A*A*.
- Tutti i fattori compresi in un effetto nidificato devono essere univoci. Di conseguenza, se *A* è un fattore, non è possibile specificare *A(A)*.
- Nessun effetto può essere nidificato all'interno di una covariata. Di conseguenza, se *A* è un fattore e *X* è una covariata, non è possibile specificare *A(X)*.

Creazione di un termine nidificato

1. Selezionare un fattore o una covariata nidificati all'interno di un altro fattore, quindi fare clic sul pulsante freccia.
2. Fare clic su **(Entro)**.
3. Selezionare il fattore entro il quale sono nidificati il fattore o la covariata precedenti, quindi fare clic sul pulsante freccia.
4. Fare clic su **Aggiungi termine**.

È anche possibile includere effetti di interazione o aggiungere più livelli di nidificazione al termine nidificato.

Effetti casuali: I fattori a effetti casuali sono campi i cui valori, contenuti nel file dei dati, possono essere considerati un campione casuale di una popolazione di valori più ampia. Possono essere utilizzati per descrivere la variabilità di eccesso dell'obiettivo. Per impostazione predefinita, se è stato selezionato più di un soggetto nella scheda Struttura dati viene creato un Blocco effetti casuali per ogni soggetto oltre quello interno. Per esempio, se sono stati selezionati i soggetti Scuola, Classe e Studente nella scheda Struttura dati, vengono creati automaticamente i seguenti blocchi di effetti casuali:

- Effetto casuale 1: il soggetto è scuola (nessun effetto, solo intercettazione)
- Effetto casuale 2: il soggetto è scuola * classe (nessun effetto, solo intercettazione)

È possibile utilizzare i blocchi di effetti casuali nei seguenti modi:

1. Per aggiungere un nuovo blocco, fare clic su **Aggiungi blocco...**. Viene visualizzata la finestra di dialogo "Blocco effetti casuali".
2. Per modificare un blocco esistente, selezionare il blocco che si desidera modificare e fare clic su **Modifica blocco...**. Viene visualizzata la finestra di dialogo "Blocco effetti casuali".
3. Per eliminare uno o più blocchi, selezionare i blocchi da eliminare e fare clic sul pulsante di eliminazione.

Blocco effetti casuali: Immettere gli effetti nel modello selezionando uno o più campi nell'elenco di sorgenti e trascinandoli nell'elenco degli effetti. Il tipo di effetto creato dipende dall'area sensibile nella quale si rilascia la selezione. I campi categoriali (flag, nominali e ordinali) vengono utilizzati come fattori nel modello e i campi continui vengono utilizzati come covariate.

- **Principale.** I campi rilasciati vengono visualizzati come effetti principali separati in fondo all'elenco degli effetti.
- **2 vie.** Tutte le possibili coppie dei campi rilasciati vengono visualizzate come interazioni a 2 vie in fondo all'elenco degli effetti.
- **3 vie.** Tutti i possibili gruppi di tre dei campi rilasciati vengono visualizzati come interazioni a 3 vie in fondo all'elenco degli effetti.
- ***.** La combinazione di tutti i campi rilasciati viene visualizzata come una singola interazione nella parte inferiore dell'elenco degli effetti.

I pulsanti a destra del generatore di effetti consentono di eseguire diverse azioni.

Tabella 11. Descrizioni dei pulsanti del generatore di effetti.

Icona	Descrizione
	Eliminare i termini dal modello selezionando i termini che si desidera eliminare e facendo clic sul pulsante di eliminazione.
	Riordinare i termini all'interno del modello selezionando i termini che si desidera riordinare e facendo clic sulla freccia rivolta verso l'alto o verso il basso.
	Aggiungere termini nidificati al modello utilizzando la finestra di dialogo "Aggiungi un termine personalizzato" a pagina 199, facendo clic sul pulsante Aggiungi un termine personalizzato.

Includi intercettazione. L'intercettazione non è inclusa nel modello degli effetti casuali per impostazione predefinita. Se è possibile presumere che i dati passino attraverso l'origine, l'intercettazione può essere esclusa.

Definisci gruppi covarianza per. I campi categorici specificati in questo punto definiscono insiemi indipendenti di parametri di covarianza degli effetti casuali; uno per ciascuna categoria definita dalla classificazione incrociata dei campi di raggruppamento. Per ogni blocco di effetti casuali è possibile specificare un insieme diverso di campi di raggruppamento. Tutti i soggetti presentano lo stesso tipo di covarianza; i soggetti all'interno dello stesso gruppo di covarianza presentano gli stessi valori per i parametri.

Combinazione soggetti. Consente di specificare soggetti degli effetti casuali da combinazioni predefinite di soggetti dalla scheda Struttura dati. Per esempio, se *Scuola*, *Classe* e *Studente* sono definiti nell'ordine come soggetti nella scheda Struttura dati, l'elenco a discesa Combinazione soggetti includerà le opzioni **Nessuna**, **Scuola**, **Scuola * Classe** e **Scuola * Classe * Studente**.

Tipo di covarianza dell'effetto casuale. Specifica la struttura di covarianza per i residui. Sono disponibili le seguenti strutture:

- Autoregressivo di primo ordine (AR1)
- Autoregressivo a media mobile (1,1) (ARMA11)
- Simmetria composta
- Diagonale
- Identità scalata
- Toeplitz
- Non strutturato
- Componenti della varianza

Peso e offset: Peso analisi. Il parametro di scala è un parametro del modello stimato correlato alla varianza della risposta. I pesi dell'analisi sono valori "noti" che possono variare a seconda delle osservazioni. Se il campo del peso dell'analisi è stato specificato, il parametro scala, che è correlato alla varianza della risposta, viene diviso per i valori del peso dell'analisi per ciascuna osservazione. Per l'analisi non vengono usati i record con valori di pesi analisi inferiori o uguali a 0 o mancanti.

Offset. Il termine offset è un predittore "strutturale". Il suo coefficiente non è stimato dal modello ma si presume che abbia il valore 1; pertanto, i valori dell'offset vengono semplicemente aggiunti al predittore lineare dell'obiettivo. Ciò è particolarmente utile nei modelli di regressione di Poisson, nei quali ogni caso può avere diversi livelli di esposizione all'evento di interesse.

Per esempio, nella modellazione dei tassi di incidente per singolo conducente esiste una differenza significativa tra un autista responsabile di un incidente in tre anni di esperienza e un autista responsabile di un incidente in 25 anni. Il numero di incidenti può essere rappresentato come una risposta di Poisson o binomiale negativa con un collegamento log se il logaritmo naturale dell'esperienza del conducente viene incluso come termine di offset.

Altre combinazioni dei tipi di distribuzione e di collegamento richiederebbero altre trasformazioni della variabile di offset.

Opzioni di creazione generali: Queste opzioni consentono di specificare alcuni criteri avanzati per la creazione del modello.

Ordinamento. Questi comandi determinano l'ordine delle categorie per l'obiettivo e i fattori (input categoriali) allo scopo di determinare l'"ultima" categoria. L'impostazione del criterio di ordinamento dell'obiettivo viene ignorata se l'obiettivo non appartiene ad una categoria o se è specificata una categoria di riferimento personalizzata nelle impostazioni “Obiettivo” a pagina 196.

Regole di arresto. È possibile specificare il numero massimo di iterazioni che verranno eseguite nell'algoritmo. L'algoritmo utilizza un doppio processo iterativo costituito da un loop interno e da uno esterno. Il valore specificato per il numero massimo di iterazioni si applica ad entrambi i loop. Specifica un intero non negativo. Il valore predefinito è 100.

Impostazioni post-stima. Queste impostazioni determinano come viene calcolata parte dell'output del modello per la visualizzazione.

- **Livello di confidenza.** Si tratta del livello di confidenza utilizzato per calcolare stime di intervallo per i coefficienti del modello. Specificare un valore maggiore di 0 e minore di 100. Il valore di default è 95.
- **Gradi di libertà.** Specifica la modalità di calcolo dei gradi di libertà per i test di significatività. Scegliere **Fissi per tutti i test (metodo dei residui)** se la dimensione campione è sufficientemente grande o i dati sono bilanciati oppure il modello utilizza un tipo di covarianza più semplice, per esempio identità scalata o diagonale. Questa è l'opzione di default. Scegliere **Diversi nei singoli test (approssimazione di Satterthwaite)** se la dimensione campione è piccola o i dati non sono bilanciati oppure il modello utilizza un tipo di covarianza complesso, per esempio non strutturato.

- **Test degli effetti fissi e dei coefficienti.** Si tratta del metodo per il calcolo della matrice di covarianza delle stime dei parametri. Scegliere la stima robusta se si teme che le ipotesi di modello vengano violate.

Stima: L'algoritmo di creazione modelli utilizza un doppio processo iterativo costituito da un loop interno e da uno esterno. Le seguenti impostazioni vengono applicate al loop interno.

Convergenza parametri.
La convergenza viene presunta se la modifica assoluta massima o la modifica relativa massima delle stime del parametro sono inferiori al valore specificato, che non deve essere negativo. Il criterio non viene utilizzato se il valore specificato è 0.

Convergenza verosimiglianza.
La convergenza viene presunta se la modifica assoluta o la modifica relativa nella funzione di verosimiglianza logaritmica sono inferiori al valore specificato, che non deve essere negativo. Il criterio non viene utilizzato se il valore specificato è 0.

Convergenza hessiana.
Per la specifica **Assoluta**, la convergenza viene presunta se la statistica basata sulla convergenza hessiana è inferiore al valore specificato. Per la specifica **Relativa**, si presume la convergenza se la statistica è minore del prodotto tra il valore specificato e il valore assoluto della log-verosimiglianza. Il criterio non viene utilizzato se il valore specificato è 0.

Max fasi di calcolo del punteggio di Fisher.
Specifica un intero non negativo. Il valore 0 specifica il metodo Newton-Raphson. I valori maggiori di 0 indicano di utilizzare l'algoritmo di calcolo del punteggio di Fisher fino al numero iterativo *n*, dove *n* è il numero intero specificato, con Newton-Raphson successivamente.

Tolleranza della singolarità
Questo valore viene usato come tolleranza nel controllo della singolarità. Specificare un valore positivo.

Nota: Per impostazione predefinita viene utilizzata la convergenza dei parametri, dove viene selezionata la massima variazione **Assoluta** rispetto alla tolleranza 1E-6. Questa impostazione potrebbe produrre risultati che differiscono dai risultati ottenuti nelle versioni precedenti alla versione 22. Per riprodurre i risultati di versioni precedenti alla versione 22, utilizzare **Relativo** per il criterio Convergenza dei parametri e conservare il valore predefinito di tolleranza 1E-6.

Generale: Nome modello. È possibile generare il nome del modello automaticamente in base ai campi obiettivo oppure specificare un nome personalizzato. Il nome generato automaticamente è il nome del campo obiettivo. Se sono presenti più obiettivi, il nome del modello è l'elenco dei nomi dei campi in ordine, collegati dalla "e" commerciale. Ad esempio se *campo1 campo2 campo3* sono obiettivi, il nome del modello è: *campo1 & campo2 & campo3*.

Rendi disponibile per il calcolo del punteggio. Quando viene calcolato il punteggio del modello, devono essere generati gli elementi selezionati in questo gruppo. Il valore previsto (per tutti gli obiettivi) e la confidenza (per gli obiettivi categoriali) vengono sempre calcolati durante il calcolo del punteggio del modello. La confidenza calcolata può essere basata sulla probabilità del valore atteso (la probabilità prevista più alta) o sulla differenza tra la probabilità prevista più alta e la seconda probabilità prevista più alta.

- **Probabilità prevista per gli obiettivi categoriali.** Genera le probabilità previste per gli obiettivi categoriali. Viene creato un campo per ogni categoria.
- **Punteggi di propensione per gli obiettivi flag.** Per i modelli con obiettivi di tipo flag (che restituiscono un risultato sì o nessuna previsione), è possibile richiedere i punteggi di propensione grezza che indicano la verosimiglianza del risultato vero specificato per il campo obiettivo. Il modello produce punteggi di propensione grezza; se le partizioni sono attive, il modello produce anche punteggi di propensione regolata basati sulla partizione di test.

Medie stimate: Questa scheda consente di visualizzare le medie marginali stimate per i livelli di fattori e le interazioni dei fattori. Le medie marginali stimate non sono disponibili per i modelli multinomiali.

Termini. Visualizza l'elenco dei termini del modello negli effetti fissi interamente composti da campi categoriali. Selezionare ogni termine per il quale si desidera che il modello produca medie marginali stimate.

- **Tipo contrasto.** Specifica il tipo di contrasto da utilizzare per i livelli del campo contrasto. Se viene selezionata l'opzione **Nessuno**, non viene prodotto alcun contrasto. **A coppie** produce confronti a coppie per tutte le combinazioni di livelli dei fattori specificati. Questo è l'unico contrasto disponibile per le interazioni tra fattori. I contrasti **Deviazione** confrontano ogni livello del fattore con la media finale. I contrasti **Semplice** confrontano ogni livello del fattore, ad eccezione dell'ultimo, con l'ultimo livello. L'"ultimo" livello è determinato dal criterio di ordinamento per i fattori specificato in Opzioni di costruzione. Si noti che tutti questi tipi di contrasto non sono ortogonali.
- **Campo contrasto.** Specifica un fattore i cui livelli vengono confrontati utilizzando il tipo di contrasto selezionato. Se è stato selezionato il tipo di contrasto **Nessuno**, non sarà possibile (o necessario) selezionare alcun campo contrasto.

Campi continui. I campi continui elencati vengono estratti dai termini negli effetti fissi che utilizzano campi continui. Nel calcolo delle medie marginali stimate, le covariate sono fisse ai valori specificati. Selezionare la media o specificare un valore personalizzato.

Visualizza medie stimate in termini di. Specifica se calcolare le medie marginali stimate in base alla scala originale dell'obiettivo o in base alla trasformazione della funzione di collegamento. **Scala obiettivo originale** calcola le medie marginali stimate per l'obiettivo. Si noti che se l'obiettivo viene specificato utilizzando l'opzione eventi/prove, vengono calcolate le medie marginali stimate per la parte eventi/prove anziché per il numero di eventi. **Trasformazione funzione collegamento** calcola le medie marginali stimate per il predittore lineare.

Adatta per confronti multipli utilizzando. Quando si eseguono test di ipotesi con più contrasti, il livello di significatività generale può essere adattato in base ai livelli di significatività dei contrasti inclusi. Ciò consente di selezionare il metodo di adattamento.

- **Differenza meno significativa.** Questo metodo non controlla la probabilità generale di rifiuto delle ipotesi che alcuni contrasti lineari siano diversi dai valori di ipotesi null.
- *Bonferroni sequenziale*. Una procedura di Bonferroni con scarti sequenzialmente decrescenti, molto meno conservativa in termini di rifiuto di singole ipotesi, ma che mantiene lo stesso livello di significatività globale.
- *Sidak sequenziale*. Una procedura di Sidak con scarti sequenzialmente decrescenti, molto meno conservativa in termini di rifiuto di singole ipotesi, ma che mantiene lo stesso livello di significatività globale.

Il metodo della differenza meno significativa è meno conservativo del metodo Sidak sequenziale che a sua volta è meno conservativo rispetto al metodo Bonferroni sequenziale. La differenza meno significativa rifiuterà almeno lo stesso numero di ipotesi individuali del Sidak sequenziale che a sua volta rifiuterà almeno lo stesso numero di ipotesi individuali del Bonferroni sequenziale.

Vista modello: Per default, viene mostrata la visualizzazione Riepilogo modello. Per un'altra visualizzazione del modello, selezionarla dalle miniature delle visualizzazioni.

Riepilogo del modello: Questa visualizzazione è una snapshot, un riepilogo del modello e del suo adattamento.

Tabella. La tabella identifica l'obiettivo, la distribuzione della probabilità e la funzione di collegamento specificata in Impostazioni obiettivo. Se l'obiettivo è definito da eventi e prove, la cella viene suddivisa in

modo da mostrare il campo eventi ed il campo prove o un numero fisso di prove. Vengono inoltre visualizzati il criterio di informazione di Akaike corretto per il campione finito (AICC) e il criterio di informazione bayesiano (BIC).

- *Akaike corretto*. Una misura per selezionare e confrontare modelli misti basata sulla verosimiglianza logaritmica -2 (ristretta). I valori più bassi indicano i modelli migliori. AICC corregge AIC in presenza di campioni piccoli. Mano mano che aumenta la dimensione del campione, il criterio AICC converge nel criterio AIC.
- *Bayesiano*. Una misura per selezionare e confrontare modelli basata sulla verosimiglianza logaritmica -2. I valori più bassi indicano i modelli migliori. Anche BIC penalizza i modelli sovraparametrizzati, ma in modo più rigoroso rispetto ad AIC.

Grafico. Se l'obiettivo è categoriale, un grafico visualizza la precisione del modello finale, che corrisponde alla percentuale delle classificazioni corrette.

Struttura dati: Questa visualizzazione offre un riepilogo della struttura di dati specificata e consente di verificare che i soggetti e le misure ripetute siano stati specificati correttamente. Per ogni campo soggetto e per ogni campo misure ripetute vengono visualizzate le informazioni osservate per il primo soggetto e l'obiettivo. Viene inoltre visualizzato il numero di livelli di ogni campo soggetto e campo misure ripetute.

Previsioni e osservazioni: Per i target continui, inclusi gli obiettivi specificati come eventi/prove, viene visualizzato un grafico a dispersione in bin dei valori previsti sull'asse verticale in base ai valori osservati sull'asse orizzontale. In teoria i punti dovrebbero trovarsi su una linea a 45 gradi; da questa visualizzazione si può capire se il modello è particolarmente carente nella previsione di determinati record.

Classificazione: Per gli obiettivi categoriali, visualizza la classificazione incrociata dei valori osservati e previsti in una mappa termica, più la percentuale globale corretta.

Stili di tabella. Sono disponibili diversi stili di visualizzazione, a cui è possibile accedere dall'elenco a discesa **Stile**.

- **Percentuali di riga.** Questa opzione visualizza la percentuale di righe (il numero di celle espresso come percentuale del totale di righe) nelle celle. Questa è l'opzione di default.
- **Conteggi delle celle.** Questa opzione visualizza il numero di celle nelle celle. L'ombreggiatura per la mappa termica è comunque basata sulle percentuali di riga.
- **Mappa termica.** Questa opzione non visualizza alcun valore nelle celle, ma solo l'ombreggiatura.
- **Compresso.** Questa opzione non visualizza alcuna intestazione di righe o colonne o valori nelle celle. Può essere utile se l'obiettivo ha molte categorie.

Mancante. Se, per alcuni record, non sono presenti valori nell'obiettivo, i valori vengono visualizzati in una riga **(Mancante)** al di sotto di tutte le righe valide. I record con valori mancanti non contribuiscono alla percentuale globale corretta.

Obiettivi multipli. Se sono presenti più obiettivi di categoria, ciascun obiettivo è visualizzato in una tabella separata ed è presente un elenco a discesa **Obiettivo** che controlla l'obiettivo da visualizzare.

Tabelle di grandi dimensioni. Se l'obiettivo visualizzato contiene più di 100 categorie, non viene visualizzata alcuna tabella.

Effetti fissi: Questa visualizzazione mostra le dimensioni di ogni effetto fisso nel modello.

Stili. Sono disponibili diversi stili di visualizzazione, a cui è possibile accedere dall'elenco a discesa **Stile**.

- **Diagramma.** Questo è un grafico in cui è effetti sono disposti dall'alto verso il basso in base all'ordine in cui sono stati specificati nelle impostazioni Effetti fissi. Le linee di collegamento del diagramma vengono pesate in base alla significatività degli effetti, con la maggiore ampiezza della linea corrispondente agli effetti più significativi (minori valori *p*). Questa è l'opzione di default.

- **Tabella.** Tabella ANOVA per gli effetti generali e specifici del modello. I singoli effetti vengono ordinati dall'alto verso il basso nell'ordine specificato nelle impostazioni Effetti fissi.

Significatività. Il dispositivo di scorrimento Significatività controlla gli effetti mostrati nella visualizzazione. Gli effetti con valori di significatività superiori al valore del dispositivo di scorrimento vengono nascosti. L'azione del dispositivo di scorrimento non modifica il modello, consente semplicemente di evidenziare gli effetti più importanti. Il valore di default è 1,00, ovvero gli effetti non vengono filtrati in base alla significatività.

Coefficienti fissi: Questa visualizzazione mostra il valore di ogni coefficiente fisso nel modello. Si noti che i fattori (predittori categoriali) sono codificati mediante un indicatore nel modello, in modo tale che agli **effetti** contenenti fattori possano essere associati più **coefficienti**, uno per ogni categoria esclusa la categoria corrispondente al coefficiente ridondante.

Stili. Sono disponibili diversi stili di visualizzazione, a cui è possibile accedere dall'elenco a discesa **Stile**.

- **Diagramma.** Questo grafico visualizza prima l'intercettazione, quindi ordina gli effetti dall'alto verso il basso nell'ordine in cui sono stati specificati nelle impostazioni Effetti fissi. Negli effetti che contengono fattori i coefficienti vengono ordinati in ordine crescente in base ai valori dei dati. Le linee di collegamento del diagramma sono colorate e pesate in base alla significatività dei coefficienti, con la maggiore ampiezza della linea corrispondente ai coefficienti più significativi (minori valori *p*). Questo è lo stile di default.
- **Tabella.** Indica i valori, i test di significatività e gli intervalli di confidenza per i singoli coefficienti del modello. Dopo l'intercettazione, gli effetti vengono ordinati dall'alto verso il basso nell'ordine specificato nelle impostazioni Effetti fissi. Negli effetti che contengono fattori i coefficienti vengono ordinati in ordine crescente in base ai valori dei dati.

Multinomiale. Se è attiva la distribuzione multinomiale, l'elenco a discesa Multinomiale controlla la categoria obiettivo da visualizzare. Il criterio di ordinamento dei valori nell'elenco è determinato dalle opzioni specificate nelle impostazioni Opzioni di costruzione.

Esponenziale. Visualizza le stime del coefficiente esponenziale e gli intervalli di confidenza per alcuni tipi di modelli, inclusi Regressione logistica binaria (distribuzione binomiale e collegamento logit), Regressione logistica nominale (distribuzione multinomiale e collegamento logit), Regressione binomiale negativa (distribuzione binomiale negativa e collegamento log) e Modello log-lineare (distribuzione di Poisson e collegamento log).

Significatività. È disponibile un dispositivo di scorrimento Significatività che controlla i coefficienti mostrati nella visualizzazione. I coefficienti con valori di significatività superiori al valore del dispositivo di scorrimento vengono nascosti. L'azione del dispositivo di scorrimento non modifica il modello, consente semplicemente di evidenziare i coefficienti più importanti. Il valore di default è 1,00, ovvero i coefficienti non vengono filtrati in base alla significatività.

Covarianze di effetti casuali: Questa visualizzazione mostra la matrice di covarianza degli effetti casuali (**G**).

Stili. Sono disponibili diversi stili di visualizzazione, a cui è possibile accedere dall'elenco a discesa **Stile**.

- **Valori di covarianza.** Questa è una mappa termica della matrice di covarianza in cui gli effetti sono ordinati dall'alto verso il basso nell'ordine con cui sono stati specificati nelle impostazioni Effetti fissi. I colori nel diagramma di correlazione corrispondono ai valori delle celle mostrati nella chiave. Questa è l'opzione di default.
- **Diagramma di correlazione.** Si tratta di una mappa termica della matrice di covarianza.
- **Compresso.** Questa è una mappa termina della matrice di covarianza senza le intestazioni di righe e colone.

Blocchi. Se sono presenti più blocchi di effetti casuali, è disponibile un elenco a discesa Blocco per la selezione del blocco da visualizzare.

Gruppi. Se un blocco di effetti casuali include una specifica di blocco, sarà disponibile un elenco a discesa Gruppo che consente di selezionare il livello di gruppo da visualizzare.

Multinomiale. Se è attiva la distribuzione multinomiale, l'elenco a discesa Multinomiale controlla la categoria obiettivo da visualizzare. Il criterio di ordinamento dei valori nell'elenco è determinato dalle opzioni specificate nelle impostazioni Opzioni di costruzione.

Parametri di covarianza: Questa visualizzazione mostra le stime dei parametri di covarianza e le statistiche associate agli effetti casuali e residui. Si tratta di risultati avanzati ma fondamentali che forniscono informazioni sull'idoneità della struttura di covarianza.

Tabella riassuntiva. Si tratta di un riferimento rapido per il numero di parametri nelle matrici di covarianza dei residui (**R**) e di effetti casuali (**G**), la classificazione (numero di colonne) nelle matrici di progetto degli effetti fissi (**X**) e degli effetti casuali (**Z**) ed il numero di oggetti definiti mediante i campi soggetto che definiscono la struttura dei dati.

Tabella Parametri di covarianza. Per l'effetto selezionato, per ciascun parametri di covarianza sono visualizzati la stima, l'errore standard e l'intervallo di confidenza. Il numero di parametri visualizzati varia a seconda della struttura di covarianza per l'effetto e, per i blocchi di effetti casuali, del numero di effetti del blocco. Se si verifica che i parametri esterni alla diagonale non sono significativi, è possibile utilizzare una struttura di covarianza più semplice.

Effetti. Se sono presenti blocchi di effetti casuali, è disponibile un elenco a discesa Effetto per la selezione del blocco di effetti casuali o residui da visualizzare. L'effetto residuo è sempre disponibile.

Gruppi. Se un blocco di effetti residui o casuali include una specifica di blocco, sarà disponibile un elenco a discesa Gruppo che consente di selezionare il livello di gruppo da visualizzare.

Multinomiale. Se è attiva la distribuzione multinomiale, l'elenco a discesa Multinomiale controlla la categoria obiettivo da visualizzare. Il criterio di ordinamento dei valori nell'elenco è determinato dalle opzioni specificate nelle impostazioni Opzioni di costruzione.

Medie stimate: effetti significativi: Si tratta dei grafici visualizzati per i 10 effetti fissi per tutti i fattori "più significativi", iniziando con le interazioni a tre vie, seguite dalle interazioni a due vie e infine dagli effetti principali. Il grafico visualizza il valore del modello stimato dell'obiettivo sull'asse verticale per ogni valore dell'effetto principale (o dell'effetto elencato per primo in un'interazione) sull'asse orizzontale; viene inoltre generata una linea distinta per ogni valore del secondo effetto elencato in un'interazione a due vie; per ogni valore nel terzo effetto elencato in un'interazione a tre vie viene generato un grafico distinto; tutti gli altri predittori rimangono costanti. È una visualizzazione utile degli effetti dei coefficienti di ciascun predittore sull'obiettivo. Si noti che se non sono presenti predittori significativi, non vengono generate medie stimate.

Confidenza. Visualizza i limiti di confidenza superiore ed inferiore per le medie marginali, utilizzando il livello di confidenza specificato come parte delle opzioni di creazione.

Medie stimate: effetti personalizzati: Si tratta di tabelle e grafici per effetti di tutti i fattori fissi richiesti dall'utente.

Stili. Sono disponibili diversi stili di visualizzazione, a cui è possibile accedere dall'elenco a discesa **Stile**.

- **Diagramma.** Lo stile visualizza un grafico a linee del valore stimato del modello del target sull'asse verticale per ciascun valore dell'effetto principale (o primo effetto elencato in un'interazione) sull'asse orizzontale; viene creata una linea separata per ciascun valore del secondo effetto elencato in

un'interazione; viene creato un grafico separato per ciascun valore del terzo effetto elencato in un'interazione a tre vie; tutti gli altri predittori vengono mantenuti costanti.

Se sono stati richiesti dei contrasti, viene visualizzato un altro grafico che consente di confrontare i livelli del campo contrasto; per le interazioni, viene visualizzato un grafico per ogni combinazione di livello degli effetti diversi dal campo contrasto. Per i contrasti **a coppie**, si tratta di un grafico di rete delle distanze, cioè una rappresentazione grafica della tabella dei confronti in cui le distanze fra i nodi della rete corrispondono a differenze tra i campioni. Le linee gialle corrispondono alle differenze statisticamente significative; le linee nere corrispondono alle differenze non significative. Se si passa il mouse sopra una linea della rete viene visualizzata una descrizione con la significatività corretta della differenza tra i nodi collegati dalla linea.

Per i contrasti di **deviazione**, viene visualizzato un grafico a barre con il valore del modello stimato dell'obiettivo sull'asse verticale e i valori del campo contrasto sull'asse orizzontale; per le interazioni, viene visualizzato un grafico per ogni combinazione di livello degli effetti diversi dal campo contrasto. Le barre mostrano la differenza tra ogni livello del campo contrasto e la media globale, rappresentata da una linea orizzontale nera.

Per i contrasti **semplici**, viene visualizzato un grafico a barre con il valore del modello stimato dell'obiettivo sull'asse verticale e i valori del campo contrasto sull'asse orizzontale; per le interazioni, viene visualizzato un grafico per ogni combinazione di livello degli effetti diversi dal campo contrasto. Le barre mostrano la differenza tra ogni livello del campo contrasto (a eccezione dell'ultimo) e l'ultimo livello, rappresentato da una linea orizzontale nera.

- **Tabella.** Questo stile visualizza una tabella del valore stimato del modello dell'obiettivo, del relativo errore standard e dell'intervallo di confidenza per ciascuna combinazione di livello dei campi nell'effetto; tutti gli altri predittori vengono mantenuti costanti.

  Se sono stati richiesti i contrasti, viene visualizzata un'altra tabella con la stima, l'errore standard, il test di significatività e l'intervallo di confidenza per ogni contrasto; per le interazioni, è presente un insieme separato di righe per ogni combinazione di livello degli effetti diversi dal campo contrasto. Viene inoltre visualizzata una tabella con i risultati dei test globali; per le interazioni, è disponibile un test globale distinto per ogni combinazione di livello degli effetti diversi dal campo contrasto.

Confidenza. Attiva o disattiva la visualizzazione dei limiti di confidenza superiore ed inferiore per le medie marginali, utilizzando il livello di confidenza specificato come parte delle opzioni di creazione.

Layout. Attiva o disattiva il layout del diagramma dei contrasti a coppie. Il layout circolare mette meno in evidenza i contrasti rispetto al layout a rete ma evita la sovrapposizione delle linee.

Impostazioni: Quando si esegue il calcolo del punteggio del modello, devono essere generati gli elementi selezionati in questa scheda. Il valore previsto (per tutti gli obiettivi) e la confidenza (per gli obiettivi categoriali) vengono sempre calcolati durante il calcolo del punteggio del modello. La confidenza calcolata può essere basata sulla probabilità del valore atteso (la probabilità prevista più alta) o sulla differenza tra la probabilità prevista più alta e la seconda probabilità prevista più alta.

- **Probabilità prevista per gli obiettivi categoriali.** Genera le probabilità previste per gli obiettivi categoriali. Viene creato un campo per ogni categoria.
- **Punteggi di propensione per gli obiettivi flag.** Per i modelli con obiettivi di tipo flag (che restituiscono un risultato sì o nessuna previsione), è possibile richiedere i punteggi di propensione grezza che indicano la verosimiglianza del risultato vero specificato per il campo obiettivo. Il modello produce punteggi di propensione grezza; se le partizioni sono attive, il modello produce anche punteggi di propensione regolata basati sulla partizione di test.

Nodo Cox

La regressione di Cox consente di creare un modello predittivo per i dati della relazione tempo-evento. Il modello crea una funzione di sopravvivenza che prevede la probabilità che l'evento di interesse si sia verificato in un dato periodo *t* per valori specifici delle variabili predittore. La forma della funzione di sopravvivenza e i coefficienti di regressione per i predittori vengono stimati in base ai soggetti osservati;

il modello può quindi essere applicato a nuovi casi che hanno misurazioni per le variabili predittore. Si noti che le informazioni provenienti da soggetti censurati, ovvero quelli che non sono interessati dall'evento nel tempo di osservazione, contribuiscono notevolmente alla stima del modello.

Esempio. Nell'ambito degli sforzi per ridurre il tasso di abbandono dei clienti, una società di telecomunicazioni desidera creare un modello del "tempo di abbandono" per determinare i fattori associati ai clienti che passano rapidamente a un servizio concorrente. A tale scopo, viene selezionato un campione casuale di clienti ed il relativo tempo trascorso come clienti (indipendentemente dal loro stato attuale di clienti) e diversi campi demografici vengono estratti dal database.

Requisiti. Sono necessari uno o più campi di input ed esattamente un campo obiettivo. Inoltre, è necessario specificare un campo Tempo di sopravvivenza nel nodo Cox. Il campo obiettivo deve essere codificato in modo che il valore "falso" indichi la sopravvivenza e il valore "vero" indichi che l'evento di interesse si è verificato; deve inoltre avere un livello di misurazione *Flag* con tipo di archiviazione come stringa o numero intero. Se necessario, è possibile convertire l'archiviazione utilizzando un nodo Riempimento o Ricava. I campi impostati su *Entrambe* o *Nessuna* verranno ignorati. È necessario che i tipi dei campi utilizzati nel modello siano completamente istanziati. Il tempo di sopravvivenza può essere un qualsiasi campo numerico.

Date & ore. I campi Data & ora non possono essere utilizzati per definire direttamente il tempo di sopravvivenza; se disponibili, i campi Data & ora devono essere utilizzati per creare un campo che contiene i tempi di sopravvivenza, in base alla differenza tra la data di immissione nello studio e la data di osservazione.

Analisi di Kaplan-Meier. È possibile eseguire la regressione di Cox senza campi di input. Questa operazione è equivalente a un'analisi di Kaplan-Meier.

Opzioni dei campi del nodo Cox

Ora di sopravvivenza. Scegliere un campo numerico (con livello di misurazione *Continuo*) per rendere il nodo eseguibile. L'ora di sopravvivenza indica la durata del record di cui si sta eseguendo la previsione. Per esempio, quando si crea un modello riferito al tempo di abbandono del cliente, in questo campo viene registrata la durata della relazione del cliente con l'organizzazione. La data in cui il cliente ha aderito o abbandonato non ha effetto sul modello; è importante solo la durata.

L'ora di sopravvivenza viene interpretata come una durata senza unità. È necessario verificare che i campi di input corrispondano all'ora di sopravvivenza. Per esempio, nel caso di uno studio per calcolare gli abbandoni in base ai mesi, come input verrebbero utilizzate le vendite mensili invece che quelle annuali. Se i propri dati hanno una data di inizio e una data di fine invece che la durata, è necessario ricodificare le date in una durata a monte del nodo Cox.

Gli altri campi di questa finestra di dialogo sono i campi normalmente utilizzati in IBM SPSS Modeler. Per ulteriori informazioni, consultare l'argomento “Opzioni dei campi dei nodi Modelli” a pagina 31.

Opzioni del modello di nodo Cox

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.

Utilizza dati partizionati. Se è definito un campo partizione, questa opzione garantisce che per la creazione del modello vengano utilizzati dati solo dalla partizione di addestramento.

Crea modelli di suddivisione. Crea un modello separato per ogni valore possibile dei campi di input che sono stati specificati come campi di suddivisione. Per ulteriori informazioni, consultare l'argomento “Creazione di modelli di suddivisione” a pagina 28.

Metodo. Per l'immissione dei predittori nel modello, sono disponibili le seguenti opzioni:

- **Invio.** Questo è il metodo di default che immette tutti i termini direttamente nel modello. Nella generazione del modello non viene eseguita alcuna selezione di campo.
- **Stepwise.** Il metodo Stepwise di selezione dei campi consente di creare il modello passo per passo. Il modello iniziale è il più semplice possibile e non presenta termini di modello (ad eccezione della costante). A ogni fase vengono valutati i termini che ancora non sono stati aggiunti al modello e, se il migliore di essi aumenta notevolmente il potere predittivo del modello, viene aggiunto. Inoltre, i termini che al momento si trovano nel modello vengono rivalutati per stabilire se è possibile rimuoverli senza pregiudicare eccessivamente il modello. Se questa possibilità esiste, i campi vengono rimossi. Il processo viene ripetuto e vengono aggiunti e/o rimossi altri termini. Quando non è più possibile aggiungere altri termini per migliorare il modello e non è più possibile rimuoverne senza danneggiare il modello, viene generato il modello finale.
- **Stepwise all'indietro.** Il metodo Stepwise All'indietro essenzialmente è l'opposto del metodo Stepwise. Con questo metodo, il modello iniziale contiene tutti i termini come predittori. Ad ogni passo, i termini nel modello vengono valutati e tutti quelli che possono essere rimossi senza pregiudicare il modello vengono eliminati. Inoltre, i termini precedentemente rimossi vengono rivalutati per stabilire se il migliore di essi aumenta notevolmente il potere predittivo del modello. In tal caso, viene nuovamente aggiunto al modello. Quando non è più possibile rimuovere altri termini per migliorare il modello e non è più possibile aggiungerne senza danneggiarlo notevolmente, viene generato il modello finale.

Nota: i metodi automatici (inclusi Stepwise e Stepwise All'indietro) sono metodi di apprendimento altamente adattabili e hanno la forte tendenza a sovradattare i dati di addestramento. Quando vengono utilizzati questi metodi, è particolarmente importante verificare la validità del modello risultante con dati nuovi o con un campione estratto per il test mediante il nodo Partizione.

Gruppi. Se si specifica un campo gruppo, il nodo calcola modelli separati per ogni categoria del campo. Può trattarsi di qualsiasi campo categoriale (flag o nominale) con tipo di archiviazione come stringa o numero intero.

Tipo di modello. Esistono due opzioni per definire i termini nel modello. I modelli **Effetti principali** includono solo i campi di input individualmente e non verificano le interazioni (effetti di moltiplicazione) tra i vari campi di input. I modelli **Personalizzato** includono solo i termini (effetti e iterazioni principali) specificati dall'utente. Quando si seleziona questa opzione, utilizzare l'elenco Termini di modello per aggiungere o rimuovere termini nel modello.

Termini di modello. Quando si crea un modello Personalizzato, è necessario specificare esplicitamente i termini nel modello. Nell'elenco è riportato l'insieme corrente di termini per il modello. I pulsanti a destra dell'elenco Termini di modello consentono di aggiungere e rimuovere i termini di modello.

- Per aggiungere termini al modello, fare clic sul pulsante *Aggiunge nuovi termini di modello*.
- Per eliminare termini, selezionare quelli desiderati e fare clic sul pulsante *Elimina i termini di modello selezionati*.

Aggiunta di termini a un modello di regressione di Cox

Quando si richiede un modello personalizzato, è possibile aggiungere termini al modello facendo clic sul pulsante *Aggiungi nuovi termini di modello* nella scheda Modello. Verrà aperta una nuova finestra di dialogo in cui è possibile specificare i termini.

Tipo di termine da aggiungere. È possibile aggiungere termini al modello in vari modi, in relazione alla selezione dei campi di input nell'elenco Campi disponibili.

- **Interazione singola.** Viene inserito il termine che rappresenta l'interazione di tutti i campi selezionati.
- **Effetti principali.** Viene inserito un termine di effetto principale (il campo stesso) per ogni campo di input selezionato.

- **Tutte le interazioni a 2 vie.** Viene inserito un termine di interazione a due vie (il prodotto dei campi di input) per ogni coppia possibile di campi di input selezionata. Per esempio, se sono stati selezionati i campi di input *A*, *B* e *C* nell'elenco Campi disponibili, questo metodo inserirà i termini *A * B*, *A * C* e *B * C*.
- **Tutte le interazioni a 3 vie.** Viene inserito un termine di interazione a tre vie (il prodotto dei campi di input) per ogni combinazione possibile di campi di input selezionata, prendendone in considerazione tre per volta. Per esempio, se sono stati selezionati i campi di input *A*, *B*, *C* e *D* nell'elenco Campi disponibili, questo metodo inserirà i termini *A * B * C*, *A * B * D*, *A * C * D* e *B * C * D*.
- **Tutte le interazioni a 4 vie.** Viene inserito un termine di interazione a quattro vie (il prodotto dei campi di input) per ogni combinazione possibile di campi di input selezionata, prendendone in considerazione quattro per volta. Per esempio, se sono stati selezionati i campi di input *A*, *B*, *C*, *D* e *E* nell'elenco Campi disponibili, questo metodo inserirà i termini *A * B * C * D*, *A * B * C * E*, *A * B * D * E*, *A * C * D * E* e *B * C * D * E*.

Campi disponibili. Elenca i campi di input disponibili da utilizzare nella creazione dei termini di modello. Si noti che l'elenco può includere campi che non sono campi di input validi. Assicurarsi quindi che tutti i termini del modello includano solo campi di input.

Anteprima. Mostra i termini che verranno aggiunti al modello se si seleziona **Inserisci**, in base ai campi selezionati e al tipo di termine selezionato in precedenza.

Inserisci. Inserisce i termini nel modello (in base alla selezione corrente di campi e del tipo di termine) e chiude la finestra di dialogo.

Opzioni avanzate del nodo Cox

Convergenza. Queste opzioni permettono di controllare i parametri per la convergenza del modello. Quando viene eseguito il modello, le impostazioni per la convergenza controllano quante volte vengono eseguiti ripetutamente i vari parametri per verificarne l'adeguatezza dell'adattamento. Quanto maggiore è la frequenza con cui vengono eseguiti i parametri, tanto più vicini saranno i risultati (ovvero, i risultati convergeranno). Per ulteriori informazioni, consultare l'argomento “Criteri di convergenza del nodo Cox”.

Output. Queste opzioni permettono di richiedere statistiche e plot aggiuntivi, inclusa la curva di sopravvivenza, che verranno visualizzate nell'output avanzato del modello generato creato dal nodo. Per ulteriori informazioni, consultare l'argomento “Opzioni di output avanzato del nodo Cox”.

Controllo. Queste opzioni permettono di controllare i criteri per l'aggiunta e la rimozione di campi con il metodo di stima Stepwise. Se il metodo selezionato è Per blocchi, il pulsante è disattivato. Per ulteriori informazioni, consultare l'argomento “Criteri di controllo del nodo Cox” a pagina 211.

Criteri di convergenza del nodo Cox

Numero massimo di iterazioni. Consente di specificare il numero massimo di iterazioni per il modello, in modo da controllare la durata della ricerca di una soluzione durante l'esecuzione della procedura.

Convergenza verosimiglianza logaritmica. Le iterazioni si interrompono se la variazione relativa della verosimiglianza è inferiore a questo valore. Il criterio non viene utilizzato se il valore specificato è 0.

Convergenza parametri. Le iterazioni si interrompono se la variazione assoluta o relativa delle stime dei parametri è inferiore a questo valore. Il criterio non viene utilizzato se il valore specificato è 0.

Opzioni di output avanzato del nodo Cox

Statistiche. È possibile ottenere statistiche per i parametri del modello, compresi gli intervalli di confidenza per exp(*B*) e la correlazione di stime. È possibile richiedere queste statistiche in corrispondenza di ogni passo oppure solo dell'ultimo.

Visualizza funzione di base. Consente di visualizzare la funzione di rischio di base e la sopravvivenza cumulativa in corrispondenza della media delle covariate.

Grafici

I grafici agevolano la valutazione del modello stimato e l'interpretazione dei risultati. È possibile tracciare le funzioni di sopravvivenza, rischio e trasformazione ln (-ln).

- *Sopravvivenza*. Visualizza la funzione di sopravvivenza cumulata in scala lineare.
- *Rischio*. Visualizza la funzione di rischio cumulato in scala lineare.
- **Logaritmo meno logaritmo.** Visualizza la stima di sopravvivenza cumulativa dopo l'applicazione della trasformazione ln(-ln) alla stima.
- *Uno meno sopravvivenza*. Visualizza il complemento a uno della funzione di sopravvivenza su una scala lineare.

Traccia una linea distinta per ciascun valore. Questa opzione è disponibile solo per i campi categoriali.

Valore da utilizzare per i plot. Poiché queste funzioni dipendono dai valori dei predittori, è necessario utilizzare valori costanti per i predittori per rappresentare graficamente le funzioni rispetto al tempo. L'impostazione di default è l'utilizzo della media di ogni predittore come valore costante, ma è possibile immettere valori personalizzati utilizzando la griglia. Per gli input categoriali viene utilizzata una codifica a indicatori, in modo che esista un coefficiente di regressione per ogni categoria tranne l'ultima. Pertanto un input categoriale ha un valore medio per ogni contrasto tra indicatori uguale alla proporzione dei casi della categoria corrispondente al contrasto tra indicatori.

Criteri di controllo del nodo Cox

Criterio di rimozione. Selezionare **Rapporto di verosimiglianza** per ottenere un modello più solido. Per diminuire i tempi richiesti per la creazione del modello, è possibile selezionare **Wald**. Esiste l'opzione aggiuntiva **Condizionale**, che consente di eseguire la verifica dell'eliminazione in base alla probabilità della statistica del rapporto di verosimiglianza basato sulle stime condizionali dei parametri.

Soglie di significatività per i criteri RL. Questa opzione permette di specificare i criteri di selezione in base alla probabilità statistica (il valore *p*) associata a ciascun campo. I campi verranno aggiunti al modello solo se il valore *p* associato è inferiore al valore di **inserimento** e verranno rimossi solo se il valore *p* è maggiore del valore di **eliminazione**. Il valore di **inserimento** deve essere inferiore al valore di **eliminazione**.

Opzioni della scheda Impostazioni per il nodo Cox

Prevedi la sopravvivenza in momenti futuri. Specificare uno o più momenti futuri. La sopravvivenza, cioè la probabilità o meno di ogni caso di sopravvivere almeno per il periodo specificato (dal momento presente) senza che si verifichi l'evento terminale, viene prevista per ogni record per ogni valore di tempo, una predizione per valore. Si noti che la sopravvivenza è il valore "falso" del campo obiettivo.

- **Intervalli regolari.** I valori del tempo di sopravvivenza vengono generati dalle impostazioni specificate per **Intervallo di tempo** e **Numero di periodi di tempo di cui calcolare il punteggio**. Se per esempio sono richiesti 3 periodi di tempo con un intervallo di 2, la sopravvivenza verrà prevista per i momenti futuri 2, 4, 6. Ogni record viene valutato agli stessi valori di tempo.
- **Campi ora.** Vengono forniti i tempi di sopravvivenza per ogni record nel campo ora scelto (una previsione per campo); ciascun record può quindi essere valutato in tempi diversi.

Tempo di sopravvivenza passato. Specificare il tempo di sopravvivenza del record fino ad ora — ad esempio, la durata di un cliente esistente come campo. Il calcolo del punteggio della probabilità di sopravvivenza in un momento futuro sarà condizionata dal tempo di sopravvivenza passato.

Nota: i valori dei tempi di sopravvivenza futuri e passati devono essere compresi nell'intervallo dei tempi di sopravvivenza nei dati utilizzati per addestrare il modello. Ai record con tempi non compresi in tale intervallo verrà assegnato un punteggio null.

Accoda tutte le probabilità. Specifica se le probabilità di ogni categoria del campo di output vengono aggiunte a ogni record elaborato dal nodo. Se questa opzione non è selezionata, verrà aggiunta solo la probabilità della categoria prevista. Le probabilità vengono calcolate per ogni momento futuro.

Calcola la funzione di rischio cumulativo. Specifica se il valore del rischio cumulativo viene aggiunto a ogni record. Il rischio cumulativo viene calcolato per ogni momento futuro.

Nugget del modello di Cox

I modelli di regressione di Cox rappresentano le equazioni stimate dai nodi Cox. Esso contiene tutte le informazioni intercettate dal modello, nonché informazioni sulla performance e la struttura del modello.

Quando viene eseguito un flusso che contiene un modello di regressione di Cox generato, il nodo aggiunge due nuovi campi contenenti la previsione del modello e la probabilità associata. I nomi dei nuovi campi derivano dal nome del campo di output di cui si sta eseguendo la previsione, a cui viene aggiunto il prefisso *$C-* per la categoria prevista e *$CP-* per la probabilità associata e il suffisso corrispondente al numero dell'intervallo di tempo futuro oppure al nome del campo ora che definisce l'intervallo di tempo. Per esempio, per un campo di output denominato *abbandono* e due intervalli di tempo futuri definiti a intervalli regolari, i nuovi campi verrebbero denominati *$C-abbandono-1*, *$CP-abbandono-1*, *$C-abbandono-2* e *$CP-abbandono-2*. Se i momenti futuri sono definiti con un campo ora *durata*, i nuovi campi sarebbero *$C-abbandono_durata* e *$CP-abbandono_durata*.

Se viene selezionata l'opzione **Accoda tutte le probabilità** nel nodo Cox, per ogni momento futuro verranno aggiunti due ulteriori campi contenenti le probabilità di sopravvivenza e non sopravvivenza per ogni record. Tali campi aggiuntivi vengono denominati in base al nome del campo di output, a cui viene aggiunto il prefisso *$CP-<valore falso>-* per la probabilità di sopravvivenza e *$CP-<valore vero>-* per la probabilità che l'evento si sia verificato e il suffisso corrispondente al numero dell'intervallo di tempo futuro. Per esempio, per un campo di output in cui il valore "falso" è 0 e il valore "vero" è 1 e due intervalli di tempo futuri definiti a intervalli regolari, i nuovi campi verrebbero denominati *$CP-0-1*, *$CP-1-1*, *$CP-0-2* e *$CP-1-2*. Se i momenti futuri sono definiti con un solo campo ora *durata*, i nuovi campi sarebbero *$CP-0-1* e *$CP-1-1*, poiché c'è un solo intervallo futuro.

Se viene selezionata l'opzione **Calcola la funzione di rischio cumulativo** nel nodo Cox, per ogni momento futuro verrà aggiunto un ulteriore campo contenente la funzione di rischio cumulativo per ogni record. Tali campi aggiuntivi vengono denominati in base al nome del campo di output, a cui viene aggiunto il prefisso *$CH-* e il suffisso corrispondente al numero dell'intervallo di tempo futuro o al nome del campo ora che definisce l'intervallo di tempo. Per esempio, per un campo di output denominato *abbandono* e due intervalli di tempo futuri definiti a intervalli regolari, i nuovi campi verrebbero denominati *$CH-abbandono-1* e *$CH-abbandono-2*. Se i momenti futuri sono definiti con un campo ora *durata*, il nuovo campo sarebbe *$CH-abbandono-1*.

Impostazioni dell'output della regressione di Cox

La scheda Impostazioni del nugget contiene gli stessi controlli della scheda Impostazioni del nodo del modello. I valori di default dei controlli del nugget sono determinati dai valori impostati nel nodo del modello. Per ulteriori informazioni, consultare l'argomento “Opzioni della scheda Impostazioni per il nodo Cox” a pagina 211.

Output delle opzioni avanzate della regressione di Cox

L'output avanzato della regressione di Cox fornisce informazioni dettagliate sul modello stimato e sulla relativa performance, inclusa la curva di sopravvivenza. La maggior parte delle informazioni contenute nell'output avanzato è piuttosto tecnica, quindi è necessaria una conoscenza approfondita della regressione di Cox per interpretare correttamente tali dati.

Capitolo 11. Modelli di cluster

I modelli di raggruppamento tramite cluster sono incentrati sull'identificazione di gruppi di record simili e sull'assegnazione di etichette ai record in base al gruppo di appartenenza. Questo processo viene completato senza una precedente conoscenza dei gruppi e delle relative caratteristiche. È infatti possibile che non sia noto addirittura il numero dei gruppi da cercare. La differenza tra i modelli di cluster ed altre tecniche di apprendimento automatico è l'assenza di un campo obiettivo o di un output predefinito per la previsione da parte del modello. Questi modelli vengono spesso definiti modelli di **apprendimento non supervisionato**, poiché non è disponibile alcuno standard esterno per la valutazione delle prestazioni del modello in termini di classificazione. Per questi modelli non esistono risposte *corrette* o *errate*. Il valore è determinato dalla capacità di acquisire gruppi significativi all'interno dei dati e di fornire descrizioni utili di tali raggruppamenti.

I metodi di cluster sono basati sulla misurazione delle distanze tra i record e tra i cluster. I record vengono assegnati ai cluster in modo da ridurre al minimo la distanza tra record appartenenti allo stesso cluster.

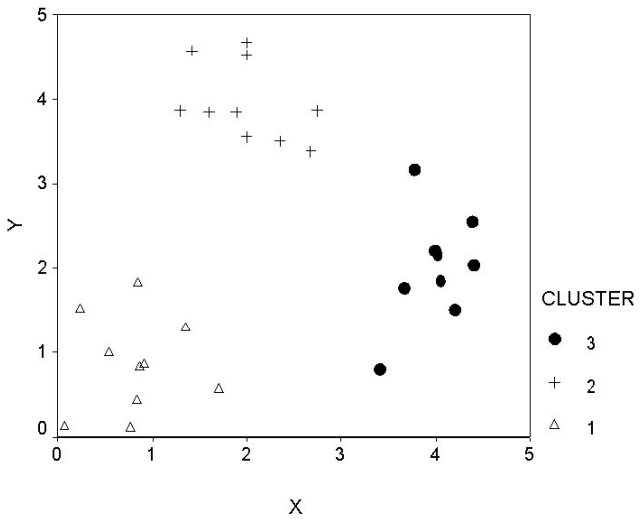


Figura 44. Modello di cluster semplice

Sono disponibili tre modelli di cluster:

Il nodo Medie K raggruppa l'insieme di dati in gruppi distinti (o cluster). Il metodo definisce un numero fisso di cluster, esegue un'assegnazione iterativa dei record ai cluster e modifica i centri di cluster finché un'ulteriore ridefinizione non consente più un miglioramento del modello. Invece di tentare di prevedere un risultato, il nodo *K*-medie utilizza un processo denominato apprendimento non supervisionato per scoprire gli schemi nell'insieme di campi di input.

Il nodo TwoStep è un metodo di raggruppamento tramite cluster in due fasi. La prima fase esegue un singolo passaggio nei dati per comprimere i dati di input non elaborati in un insieme gestibile di cluster secondari. Nella seconda fase viene utilizzato un metodo di raggruppamento tramite cluster gerarchico per unire progressivamente i cluster secondari in cluster sempre più grandi. Il nodo TwoStep offre il vantaggio di stimare automaticamente il numero ottimale di cluster per i dati di addestramento. Può gestire in modo efficiente tipi di campo misti e insiemi di dati di grandi dimensioni.

Il nodo Kohonen genera un tipo di rete neurale che può essere utilizzato per raggruppare l'insieme di dati in gruppi distinti. Al termine dell'apprendimento della rete, i record analoghi dovranno essere vicini nella mappa di output, mentre i record diversi saranno a notevole distanza. Per identificare le unità forti, è possibile controllare il numero di osservazioni catturate da ciascuna unità nel nugget del modello. In questo modo è possibile avere un'idea del numero appropriato di cluster.

I modelli di cluster vengono spesso utilizzati per creare cluster o segmenti che vengono quindi impiegati come input in analisi successive. Un esempio comune è costituito dai segmenti di mercato utilizzati dagli esperti di marketing per suddividere il mercato complessivo in sottogruppi omogenei. Ogni segmento presenta caratteristiche specifiche che influiscono sul successo delle iniziative di marketing ad esso rivolte. Se si utilizza il data mining per ottimizzare una strategia di marketing, è in genere possibile migliorare significativamente il modello utilizzato identificando i segmenti appropriati e utilizzando le relative informazioni nei modelli predittivi.

Nodo Kohonen

La rete Kohonen è un tipo di rete neurale che esegue il raggruppamento tramite cluster, conosciuta anche come **knet** o **mappa auto-organizzante**. Questo tipo di rete può essere utilizzato per raggruppare l'insieme di dati in gruppi distinti quando non si è in grado di definire immediatamente le caratteristiche di tali gruppi. I record vengono raggruppati in modo che quelli simili si trovino nello stesso gruppo o cluster e quelli dissimili in gruppi diversi.

Le unità di base sono i **neuroni**, organizzati in due strati: lo **strato di input** e lo **strato di output** (detto anche **mappa di output**). Tutti i neuroni di input vengono connessi a tutti i neuroni di output e a queste connessioni vengono associati **pesi** o **intensità**. Durante l'addestramento, le unità competono tra di loro per "aggiudicarsi" ciascun record.

La mappa di output è costituita da una griglia bidimensionale di neuroni, senza connessioni tra le unità.

I dati di input vengono presentati nello strato iniziale e i valori vengono propagati nello strato di output. Il neurone di output che fornisce la risposta di maggiore intensità viene definito **vincitore** e rappresenta la risposta per l'input.

Inizialmente tutti i pesi sono casuali. Quando un'unità si aggiudica un record, i suoi pesi (e quelli di altre unità vicine, chiamate collettivamente **vicinanza**) vengono adeguati per meglio corrispondere allo schema dei valori predittori per quel record. Vengono visualizzati tutti i record di input e i pesi vengono aggiornati di conseguenza. Questo processo viene ripetuto numerose volte fino a ottenere variazioni estremamente ridotte. Man mano che l'addestramento procede, i pesi sulle unità della griglia vengono adeguati in modo da formare una "mappa" bidimensionale dei cluster, da qui il termine **mappa auto-organizzante**.

Al termine dell'apprendimento della rete, i record analoghi dovranno essere vicini nella mappa di output, mentre i record molto diversi saranno a notevole distanza.

Contrariamente alla maggior parte dei metodi di apprendimento in IBM SPSS Modeler, le reti Kohonen *non* utilizzano un campo obiettivo. Questo tipo di apprendimento, privo di campi obiettivo, viene chiamato **apprendimento non supervisionato**. Invece di tentare di prevedere un risultato, le reti Kohonen cercano di scoprire gli schemi presenti nell'insieme di campi di input. In genere, una rete Kohonen finisce per avere poche unità che riassumono molte osservazioni (unità **forti**) e molte unità che non corrispondono a nessuna delle osservazioni (unità **deboli**). Le unità forti (e a volte altre unità adiacenti ad esse nella griglia) rappresentano i centri di cluster probabili.

Le reti Kohonen sono utilizzate anche per eseguire una **riduzione della dimensione**. La caratteristica spaziale della griglia bidimensionale fornisce una mappatura dai predittori *k* originali a due funzioni derivate che preservano le relazioni di similarità nei predittori originali. In alcuni casi, questa funzione offre lo stesso vantaggio dell'analisi fattoriale o PCA.

Si noti che il metodo per calcolare la dimensione di default della griglia di output è cambiato rispetto alle versioni precedenti di IBM SPSS Modeler. Il nuovo metodo produce in genere strati di output più piccoli, ma più rapidi da addestrare e meglio generalizzabili. Se la dimensione di default dà scarsi risultati, provare ad aumentare la dimensione della griglia di output nella scheda Livello avanzato. Per ulteriori informazioni, consultare l'argomento "Opzioni avanzate del nodo Kohonen" a pagina 216.

Requisiti. Per addestrare una rete Kohonen, è necessario avere uno o più campi con il ruolo impostato su *Input*. I campi con ruolo impostato su *Obiettivo*, *Entrambe* o *Nessuna* verranno ignorati.

Efficacia. Per generare un modello di rete Kohonen non è necessario che vi siano dati sull'appartenenza ai gruppi. Non occorre nemmeno sapere il numero di gruppi da cercare. Le reti Kohonen partono con un grande numero di unità che, nel corso dell'addestramento, gravitano verso i cluster naturali nei dati. Per identificare le unità forti e avere così un'idea del numero appropriato di cluster, controllare il numero di osservazioni catturate da ciascuna unità nel nugget del modello.

Opzioni del modello di nodo Kohonen

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.

Utilizza dati partizionati. Se è definito un campo partizione, questa opzione garantisce che per la creazione del modello vengano utilizzati dati solo dalla partizione di addestramento.

Addestramento continuo modello esistente. Per default, a ogni esecuzione di un nodo Kohonen viene creata una rete completamente nuova. Se si seleziona questa opzione, l'addestramento continuerà con l'ultima rete creata correttamente dal nodo.

Mostra grafico feedback. Se questa opzione è selezionata, durante l'addestramento viene visualizzata una rappresentazione grafica dell'array bidimensionale. La forza di ciascun nodo è rappresentata dal colore. Il rosso denota un'unità che si sta aggiudicando molti record (unità **forte**), il bianco denota un'unità che si sta aggiudicando pochi record o non se ne sta aggiudicando affatto (unità **debole**). Il feedback potrebbe non essere visualizzato se il tempo utilizzato per creare il modello è relativamente breve. Si noti che questa funzione può rallentare l'addestramento. Per accelerarlo, deselezionare questa opzione.

Condizione arresto. Il criterio di arresto di default interrompe l'addestramento, in base ai parametri interni. Come criterio di arresto è anche possibile specificare il tempo. Immettere il tempo (in minuti) di addestramento della rete.

Imposta seed random. Se non viene impostato alcun seme random, la sequenza di valori casuali utilizzata per inizializzare i pesi della rete cambierà ad ogni esecuzione del nodo. Il nodo potrebbe quindi creare modelli diversi alle varie esecuzioni, anche se le impostazioni del nodo e i valori dei dati sono esattamente gli stessi. Se si seleziona questa opzione, sarà possibile impostare il seme random su un valore specifico in modo che il modello risultante possa essere ricreato fedelmente. Un seme random specifico genera sempre la stessa sequenza di valori casuali, di conseguenza l'esecuzione del nodo produrrà sempre lo stesso modello generato.

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.

Nota: se si desidera includere campi (insiemi) nominali nel proprio modello ma si verificano problemi di memoria durante la creazione del modello o se la creazione del modello richiede troppo tempo, prendere in considerazione la ricodifica dei campi insieme di grandi dimensioni per ridurre il numero di valori oppure l'utilizzo di un campo differente con un numero minore di lavori come proxy per l'insieme di grandi dimensioni. Se, ad esempio, esiste un problema con un campo *id_prodotto* che contiene i valori per i singoli prodotti, è possibile rimuoverlo dal modello e sostituirlo con un campo *categoria_prodotto* meno dettagliato.

Ottimizza. Selezionare le opzioni per l'incremento delle prestazioni durante la creazione del modello in base alle specifiche esigenze.

- Selezionare **Velocità** per indicare all'algoritmo di non utilizzare mai il riversamento su disco per migliorare le prestazioni.
- Selezionare **Memoria** per indicare all'algoritmo di utilizzare il riversamento su disco quando necessario anche se ciò può comportare una diminuzione delle prestazioni. Questa opzione è selezionata per default.

  Nota: durante l'esecuzione in modalità distribuita, questa impostazione può essere sovrascritta dalle opzioni dell'amministratore specificate in *options.cfg*.

Accoda etichetta cluster. Selezionata per default per i nuovi modelli ma deselezionata per i modelli caricati a partire da precedenti versioni di IBM SPSS Modeler, crea un unico campo punteggio categoriale dello stesso tipo di quelli creati dai nodi Medie K e TwoStep. Questo campo stringa viene utilizzato nel nodo Cluster automatico durante il calcolo delle misure di classificazione per i diversi tipi di modelli. Per ulteriori informazioni, consultare l'argomento "Nodo Cluster automatico" a pagina 74.

Opzioni avanzate del nodo Kohonen

Le opzioni avanzate permettono agli esperti di reti Kohonen di mettere a punto il processo di addestramento. Per accedere alle opzioni avanzate, impostare la modalità su **Livello avanzato** nella scheda Livello avanzato.

Larghezza e Lunghezza. Specificare la dimensione (larghezza e lunghezza) della mappa di output bidimensionale come numero di unità di output lungo ciascuna dimensione.

Decadimento del tasso di apprendimento. Selezionare il decadimento del tasso di apprendimento lineare o esponenziale. Il **tasso di apprendimento** è un fattore di ponderazione che diminuisce con il tempo, pertanto tramite la rete vengono innanzitutto codificate le funzioni su larga scala dei dati e quindi gradatamente i dettagli più precisi.

Fase 1 e Fase 2. L'addestramento della rete Kohonen è suddiviso in due fasi. La Fase 1 è una fase di stima approssimativa, utilizzata per catturare gli schemi di massima nei dati. La Fase 2 è una fase di messa a punto, utilizzata per adeguare la mappa in modo da creare modelli in base alle caratteristiche più specifiche dei dati. Per ogni fase esistono tre parametri:

- **Vicinanza.** Imposta la dimensione iniziale (raggio) della vicinanza. Questa impostazione determina il numero di unità "vicine" che vengono aggiornate insieme all'unità vincente durante l'addestramento. Durante la fase 1, la dimensione della vicinanza parte da *Vicinanza Fase 1* e scende fino a (*Vicinanza Fase 2* + 1). Durante la fase 2, la dimensione della vicinanza parte da *Vicinanza Fase 2* e scende fino a 1,0. *Vicinanza Fase 1* deve essere maggiore di *Vicinanza Fase 2*.
- **Eta iniziale.** Imposta il valore iniziale per il tasso di apprendimento **eta**. Durante la fase 1, eta parte da *Eta iniziale Fase 1* e scende fino a *Eta iniziale Fase 2*. Durante la fase 2, eta parte da *Eta iniziale Fase 2* e scende fino a 0. *Eta iniziale Fase 1* deve essere maggiore di *Eta iniziale Fase 2*.
- **Cicli.** Imposta il numero di cicli per ogni fase di addestramento. Ogni fase continua per il numero specificato di passaggi nei dati.

Nugget del modello Kohonen

I nugget del modello Kohonen contengono tutte le informazioni intercettate dalla rete Kohonen addestrata, nonché le informazioni sull'architettura della rete.

Quando si esegue un flusso contenente un nugget del modello Kohonen, il nodo aggiunge due nuovi campi contenenti le coordinate *X* e *Y* dell'unità nella griglia di output di Kohonen che soddisfa il record completamente. I nomi dei nuovi campi derivano dal nome del modello, con l'aggiunta del prefisso *$KX-* e *$KY-*. Per esempio, se il modello è denominato *Kohonen*, i nuovi campi si chiameranno *$KX-Kohonen* e *$KY-Kohonen*.

Per comprendere meglio il risultato delle operazioni di codifica della rete Kohonen, fare clic sulla scheda Modello nel browser del nugget del modello. Verrà aperto il Visualizzatore cluster, in cui è disponibile una rappresentazione grafica dei cluster, dei campi e dei livelli di importanza. Per ulteriori informazioni, consultare l'argomento "Visualizzatore cluster - Scheda Modello" a pagina 222.

Se si preferisce visualizzare i cluster in una griglia, è possibile ottenere i risultati della rete Kohonen tramite la rappresentazione grafica dei campi *$KX-* e *$KY-* utilizzando un nodo Plot. È possibile selezionare **Agitazione X** e **Agitazione Y** nel nodo Plot per impedire che i record di ogni unità vengano rappresentati sovrapposti. Nel nodo Plot, è anche possibile sovrapporre un campo simbolico per comprendere come la rete Kohonen abbia raggruppato i dati tramite cluster.

Un'altra potente tecnica per la comprensione della rete Kohonen è l'utilizzo dell'induzione di regole per scoprire le caratteristiche che differenziano i cluster rilevati dalla rete. Per ulteriori informazioni, consultare l'argomento "nodo C5.0" a pagina 105.

Per informazioni generali relative all'utilizzo del browser del modello, consultare "Esplorazione dei nugget del modello" a pagina 42

Riepilogo del modello Kohonen

Nella scheda Riepilogo di un nugget del modello Kohonen sono visualizzate informazioni sull'architettura o topologia della rete. La lunghezza e la larghezza della mappa bidimensionale delle caratteristiche Kohonen (strato di output) vengono visualizzate come **$KX-** *model_name* e **$KY-** *model_name*. Per ogni strato di input e output, viene elencato il numero di unità contenute nello strato.

Nodo Medie K

Il nodo Medie K fornisce un metodo di **analisi dei cluster**. Tale metodo può essere utilizzato per raggruppare l'insieme di dati in gruppi distinti quando non si è in grado di definire immediatamente le caratteristiche di tali gruppi. Contrariamente alla maggior parte dei metodi di apprendimento in IBM SPSS Modeler, i modelli Medie K *non* utilizzano un campo obiettivo. Questo tipo di apprendimento, privo di campi obiettivo, viene chiamato **apprendimento non supervisionato**. Invece di cercare di prevedere un risultato, Medie K tenta di scoprire gli schemi nell'insieme di campi di input. I record vengono raggruppati in modo che quelli simili si trovino nello stesso gruppo o cluster e quelli dissimili in gruppi diversi.

Il nodo Medie K definisce un insieme di centri di cluster iniziali derivati dai dati. Quindi assegna ciascun record al cluster che gli è più simile, in base ai valori del campo di input del record. Una volta che tutti i casi sono stati assegnati, i centri di cluster vengono aggiornati in modo da riflettere il nuovo insieme di record assegnato a ciascun cluster. I record vengono quindi controllati per vedere se devono essere riassegnati a un altro cluster e il processo di iterazione dell'assegnazione di record/cluster continua finché non viene raggiunto il numero massimo di iterazioni oppure si interrompe quando il passaggio da un'iterazione all'altra non supera un limite specificato.

Nota: il modello risultante dipende in parte dall'ordine dei dati di addestramento. Pertanto, se l'ordine dei dati viene modificato e il modello viene ricreato, è possibile che si ottenga un modello di cluster finale diverso.

Requisiti. Per addestrare un modello Medie K, è necessario avere uno o più campi con il ruolo impostato su *Input*. I campi con ruolo impostato su *Output*, *Entrambi* o *Nessuno* verranno ignorati.

Efficacia. Per generare un modello Medie K non è necessario che vi siano dati sull'appartenenza ai gruppi. Il modello Medie K spesso è il metodo più veloce di raggruppare insiemi di dati di grandi dimensioni.

Opzioni del modello di nodo Medie K

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.

Utilizza dati partizionati. Se è definito un campo partizione, questa opzione garantisce che per la creazione del modello vengano utilizzati dati solo dalla partizione di addestramento.

Numero specificato di cluster. Specificare il numero di cluster da generare. Il valore predefinito è 5.

Genera campo distanza. Se questa opzione è selezionata, il nugget del modello includerà un campo contenente la distanza di ciascun record dal centro di cluster che gli è stato assegnato.

Etichetta cluster. Specificare il formato per i valori nel campo di appartenenza al cluster generato. L'appartenenza al cluster può essere indicata come **Stringa** con specificato il **prefisso Etichetta** (per esempio `"Cluster 1"`, `"Cluster 2"` e così via) o come **Numero**.

Nota: se si desidera includere campi (insiemi) nominali nel proprio modello ma si verificano problemi di memoria durante la creazione del modello o se la creazione del modello richiede troppo tempo, prendere in considerazione la ricodifica dei campi insieme di grandi dimensioni per ridurre il numero di valori oppure l'utilizzo di un campo differente con un numero minore di valori come proxy per l'insieme di grandi dimensioni. Se, ad esempio, esiste un problema con un campo *id_prodotto* che contiene i valori per i singoli prodotti, è possibile rimuoverlo dal modello e sostituirlo con un campo *categoria_prodotto* meno dettagliato.

Ottimizza. Selezionare le opzioni per l'incremento delle prestazioni durante la creazione del modello in base alle specifiche esigenze.

- Selezionare **Velocità** per indicare all'algoritmo di non utilizzare mai il riversamento su disco per migliorare le prestazioni.
- Selezionare **Memoria** per indicare all'algoritmo di utilizzare il riversamento su disco quando necessario anche se ciò può comportare una diminuzione delle prestazioni. Questa opzione è selezionata per default.

  Nota: durante l'esecuzione in modalità distribuita, questa impostazione può essere sovrascritta dalle opzioni dell'amministratore specificate in *options.cfg*.

Opzioni avanzate del nodo Medie K

Le opzioni avanzate permettono agli esperti del metodo cluster *K*-medie di mettere a punto il processo di addestramento. Per accedere alle opzioni avanzate, impostare la modalità su **Livello avanzato** nella scheda Livello avanzato.

Condizione arresto. Specificare il criterio di arresto da utilizzare nell'addestramento del modello. Il criterio di arresto **Default** corrisponde a 20 iterazioni o a una variazione < 0,000001, a seconda di quale condizione si verifica per prima. Selezionare **Personalizzato** per specificare un criterio di arresto personalizzato.

- **Massimo numero di iterazioni.** Questa opzione permette di interrompere l'addestramento del modello dopo il numero di iterazioni specificato.
- **Tolleranza del cambiamento.** Questa opzione permette di interrompere l'addestramento del modello quando la variazione maggiore nei centri di cluster per un'iterazione è inferiore al livello specificato.

Valore di codifica per insiemi. Specificare un valore tra 0 e 1,0 da utilizzare per ricodificare i campi insieme come gruppi di campi numerici. Il valore di default è la radice quadrata di 0,5 (circa 0,707107), che garantisce la ponderazione corretta dei campi flag ricodificati. Valori più vicini a 1.0 assegnano più peso ai campi insieme che ai campi numerici.

Nugget del modello Medie K

I nugget del modello Medie K contengono tutte le informazioni intercettate dal modello di cluster, nonché le informazioni sui dati di addestramento e sull'elaborazione della stima.

Quando si esegue un flusso che contiene un nodo Modelli Medie K, il nodo aggiunge due nuovi campi contenenti la classe di appartenenza e la distanza dal centro di cluster assegnato relativo a tale record. I nomi dei nuovi campi derivano dal nome del modello, con l'aggiunta del prefisso *$KM-* per la classe di appartenenza e *$KMD-* per la distanza dal centro di cluster. Per esempio, se il modello è denominato *Kmeans*, i nuovi campi si chiameranno *$KM-Kmeans* e *$KMD-Kmeans*.

Una potente tecnica utile per comprendere completamente il modello Medie K consiste nell'utilizzo dell'induzione di regole per scoprire le caratteristiche che distinguono i cluster rilevati dal modello. Per ulteriori informazioni, consultare l'argomento “nodo C5.0” a pagina 105. È anche possibile fare clic sulla scheda Modello nel browser del nugget del modello per visualizzare il visualizzatore di cluster, che fornisce una rappresentazione grafica di cluster, campi e livelli di importanza. Per ulteriori informazioni, consultare l'argomento “Visualizzatore cluster - Scheda Modello” a pagina 222.

Per informazioni generali relative all'utilizzo del browser del modello, consultare “Esplorazione dei nugget del modello” a pagina 42

Riepilogo del modello Medie K

La scheda Riepilogo di un nugget del modello Medie K contiene informazioni sui dati di addestramento, il processo di stima e i cluster definiti dal modello. Viene visualizzato il numero di cluster e la cronologia delle iterazioni. Se è stato eseguito un nodo Analisi collegato a questo nodo Modelli, in questa sezione verranno visualizzate anche le informazioni sui risultati di tale analisi.

Nodo Cluster TwoStep

Il nodo Cluster TwoStep offre una forma di **analisi dei cluster**. Tale metodo può essere utilizzato per raggruppare l'insieme di dati in gruppi distinti quando non si è in grado di definire immediatamente le caratteristiche di tali gruppi. Come nel caso dei nodi Kohonen e Medie K, i modelli Cluster TwoStep *non* utilizzano un campo obiettivo. Invece di tentare di prevedere un risultato, Cluster TwoStep cerca di scoprire gli schemi nell'insieme di campi di input. I record vengono raggruppati in modo che quelli simili si trovino nello stesso gruppo o cluster e quelli dissimili in gruppi diversi.

Cluster TwoStep è un metodo di raggruppamento tramite cluster in due fasi. La prima fase esegue un singolo passaggio nei dati, durante il quale comprime i dati di input non elaborati in un insieme gestibile di cluster secondari. La seconda fase utilizza un metodo di raggruppamento tramite cluster gerarchico per unire progressivamente i cluster secondari in cluster sempre più grandi, senza che sia richiesto un altro passaggio nei dati. Il raggruppamento tramite cluster gerarchico ha il vantaggio di non richiedere la

selezione anticipata del numero di cluster. Molti metodi di raggruppamento tramite cluster gerarchico prendono i singoli record come cluster iniziali e li uniscono in modo ricorsivo per produrre cluster sempre più grandi. Sebbene questi approcci spesso siano inefficienti nel caso di grandi quantità di dati, il raggruppamento preliminare iniziale tramite cluster del metodo TwoStep rende rapido il raggruppamento gerarchico anche per grandi insiemi di dati.

Nota: il modello risultante dipende in parte dall'ordine dei dati di addestramento. Pertanto, se l'ordine dei dati viene modificato e il modello viene ricreato, è possibile che si ottenga un modello di cluster finale diverso.

Requisiti. Per addestrare un modello Cluster TwoStep, è necessario avere uno o più campi con il ruolo impostato su *Input*. I campi con ruolo impostato su *Obiettivo*, *Entrambe* o *Nessuna* verranno ignorati. L'algoritmo Cluster TwoStep non gestisce i valori mancanti. Quando si genera il modello, i record con valori vuoti per uno qualsiasi dei campi di input verranno ignorati.

Efficacia. Il Cluster TwoStep può gestire in modo efficiente tipi di campo misti e insiemi di dati di grandi dimensioni. Inoltre è in grado di verificare diverse soluzioni di cluster e di scegliere la migliore, pertanto non è necessario sapere prima quanti cluster chiedere. Il Cluster TwoStep può essere impostato per escludere automaticamente i **valori anomali**, o casi estremamente insoliti che possono contaminare i risultati.

Opzioni del modello di nodo Cluster TwoStep

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.

Utilizza dati partizionati. Se è definito un campo partizione, questa opzione garantisce che per la creazione del modello vengano utilizzati dati solo dalla partizione di addestramento.

Standardizza campi numerici. Per default, TwoStep standardizzerà tutti i campi di input numerici sulla stessa scala, con una media di 0 e una varianza di 1. Per mantenere la graduazione originale per i campi numerici, deselezionare questa opzione. I campi simbolici non sono influenzati.

Escludi valori anomali. Se si seleziona questa opzione, i record che non risultano adatti a un cluster di base verranno automaticamente esclusi dall'analisi. Questa esclusione impedisce la distorsione dei risultati.

Il rilevamento dei valori anomali viene eseguito durante la fase preliminare di raggruppamento in cluster. Quando l'opzione è selezionata, i cluster secondari con pochi record relativi ad altri cluster secondari vengono considerati potenziali valori anomali e tre dei cluster secondari vengono ricreati escludendo tali record. L'opzione **Percentuale** controlla la dimensione al di sotto della quale si ritiene che i cluster secondari contengano potenziali valori anomali. Alcuni potenziali record di valori anomali possono essere aggiunti ai cluster secondari ricreati se sono sufficientemente simili a uno dei nuovi profili di cluster secondari. Gli altri potenziali valori anomali che non possono essere uniti vengono considerati valori anomali e vengono aggiunti a un cluster "rumore" ed esclusi dalla fase di raggruppamento in cluster gerarchici.

Quando si *determinano punteggi* dei dati con un modello TwoStep che utilizza la gestione dei valori anomali, i nuovi casi che superano una certa distanza dalla soglia (in base alla log-verosimiglianza) rispetto al più vicino cluster di base vengono considerati valori anomali e assegnati al cluster "rumore" con il nome -1.

Etichetta cluster. Specificare il formato per il campo di appartenenza al cluster generato. L'appartenenza al cluster può essere indicata come **Stringa** con specificato il **prefisso Etichetta** (per esempio `"Cluster 1"`, `"Cluster 2"` e così via) o come **Numero**.

Calcola automaticamente il numero di cluster. Il cluster TwoStep è in grado di analizzare rapidamente un gran numero di soluzioni di cluster per scegliere il numero di cluster ottimale per i dati di addestramento. Specificare un intervallo di soluzioni da provare impostando il numero **Massimo** ed il numero **Minimo** di cluster. TwoStep utilizza un processo in due fasi per determinare il numero ottimale di cluster. Nella prima fase viene selezionato il limite superiore del numero di cluster nel modello, in base alla variazione nel criterio di informazione di Bayes (BIC, Bayes Information Criterion) man mano che vengono aggiunti altri cluster. Nella seconda fase viene trovata la variazione nella distanza minima tra i cluster, per tutti i modelli con un numero di cluster inferiore alla soluzione BIC minima. La variazione maggiore nella distanza viene utilizzata per identificare il modello di cluster finale.

Specifica numero di cluster Se si conosce il numero di cluster da includere nel modello, selezionare questa opzione e specificare il numero di cluster.

Misura della distanza. Questa selezione determina la modalità di calcolo della similarità tra due cluster.

- **Verosimiglianza logaritmica.** La misura di verosimiglianza applica una distribuzione per probabilità alle variabili. Si suppone che le variabili continue vengano distribuite normalmente, mentre le variabili categoriali in base al modello multinomiale. Si suppone che tutte le variabili siano indipendenti.
- **Euclidea** La misura euclidea è la distanza in "linea retta" tra due cluster. Può essere utilizzata solo quando tutte le variabili sono continue.

Criteri di raggruppamento. Questa selezione determina la modalità di definizione del numero dei cluster mediante l'algoritmo di raggruppamento automatico. È possibile specificare il modello Criterio bayesiano di Schwarz (BIC, Bayesian Information Criterion) o il modello Criterio di informazione di Akaike (AIC, Akaike Information Criterion).

Nugget del modello Cluster TwoStep

I nugget del modello Cluster TwoStep contengono tutte le informazioni intercettate dal modello di cluster, nonché le informazioni sui dati di addestramento e sull'elaborazione della stima.

Quando viene eseguito un flusso contenente un nugget del modello Cluster TwoStep, il nodo aggiunge un nuovo campo contenente la classe di appartenenza relativa a tale record. Il nome del nuovo campo deriva dal nome del modello, con l'aggiunta del prefisso *$T-*. Per esempio, se il modello si chiama *TwoStep*, il nuovo campo verrà denominato *$T-TwoStep*.

Una potente tecnica utile per comprendere completamente il modello TwoStep consiste nell'utilizzo dell'induzione di regole per scoprire le caratteristiche che distinguono i cluster rilevati dal modello. Per ulteriori informazioni, consultare l'argomento “nodo C5.0” a pagina 105. È anche possibile fare clic sulla scheda Modello nel browser del nugget del modello per visualizzare il visualizzatore di cluster, che fornisce una rappresentazione grafica di cluster, campi e livelli di importanza. Per ulteriori informazioni, consultare l'argomento “Visualizzatore cluster - Scheda Modello” a pagina 222.

Per informazioni generali relative all'utilizzo del browser del modello, consultare “Esplorazione dei nugget del modello” a pagina 42

Riepilogo del modello TwoStep

Nella scheda Riepilogo di un nugget del modello Cluster TwoStep sono visualizzati il numero di cluster rilevati, nonché informazioni sui dati di addestramento, sull'elaborazione della stima e sulle impostazioni di creazione utilizzate.

Per ulteriori informazioni, consultare l'argomento “Esplorazione dei nugget del modello” a pagina 42.

Il Visualizzatore cluster

I modelli di cluster vengono solitamente utilizzati per cercare gruppi (o cluster) di record simili in base alle variabili esaminate, in cui la similarità tra i membri dello stesso gruppo è elevata mentre quella tra i membri di gruppi diversi è bassa. È possibile utilizzare i risultati per identificare quelle associazioni che altrimenti non sarebbero evidenti. Ad esempio, attraverso l'analisi dei cluster delle preferenze, del livello di reddito e delle abitudini di spesa dei clienti, è possibile identificare quei tipi di consumatori che con maggiore probabilità risponderanno favorevolmente a una determinata campagna di marketing.

Sono possibili due approcci all'interpretazione dei risultati in una visualizzazione cluster:

- L'esame di un cluster per determinarne le caratteristiche esclusive. *Il cluster contiene i titolari di prestiti con reddito elevato? Il cluster contiene più record degli altri?*
- L'esame dei campi di tutti i cluster per determinare il modo in cui i valori sono distribuiti tra i cluster. *Il livello di istruzione di un soggetto ne determina l'appartenenza a un cluster? Un indice di affidabilità creditizia elevato determina l'appartenenza a un cluster o a un altro?*

Mediante le visualizzazioni principali e le diverse visualizzazioni collegate nel Visualizzatore cluster, è possibile ottenere maggiori informazioni per rispondere a questi quesiti.

I seguenti nugget del modello di cluster possono essere generati in IBM SPSS Modeler:

- Nugget del modello con reti di Kohonen
- Nugget del modello Medie K
- Nugget del modello di cluster TwoStep

Per visualizzare le informazioni sui nugget del modello di cluster, fare clic con il pulsante destro del mouse sul nodo di un modello e scegliere **Sfoglia** dal menu di scelta rapida (o **Modifica** per i nodi di un flusso). In alternativa, se si sta utilizzando il nodo per la creazione dei modelli Cluster automatico, fare doppio clic sull'insieme di cluster all'interno del nugget del modello Cluster automatico. Per ulteriori informazioni, consultare l'argomento "Nodo Cluster automatico" a pagina 74.

Visualizzatore cluster - Scheda Modello

La scheda Modello dei modelli di cluster mostra una visualizzazione grafica di statistiche riassuntive e distribuzioni dei campi tra i cluster, nota come **Visualizzatore cluster**.

Nota: la scheda Modello non è disponibile per i modelli creati nelle versioni di IBM SPSS Modeler precedenti alla versione 13.

Il Visualizzatore cluster è composto da due riquadri, la visualizzazione principale a sinistra e quella collegata, o ausiliaria, a destra. Le visualizzazioni principali sono due:

- Riepilogo del modello (visualizzazione predefinita). Per ulteriori informazioni, consultare l'argomento "Visualizzazione Riepilogo del modello" a pagina 223.
- Raggruppamenti. Per ulteriori informazioni, consultare l'argomento "Visualizzazione cluster" a pagina 223.

Le visualizzazioni collegate/ausiliarie sono quattro:

- Importanza predittore. Per ulteriori informazioni, consultare l'argomento "Visualizzazione Importanza predittore nei cluster" a pagina 225.
- Dimensioni cluster (visualizzazione predefinita). Per ulteriori informazioni, consultare l'argomento "Visualizzazione Dimensioni dei cluster" a pagina 225.
- Distribuzione delle celle. Per ulteriori informazioni, consultare l'argomento "Visualizzazione Distribuzione delle celle" a pagina 225.
- Confronto tra cluster. Per ulteriori informazioni, consultare l'argomento "Visualizzazione Confronto tra cluster" a pagina 225.

Visualizzazione Riepilogo del modello

La visualizzazione Riepilogo del modello mostra una snapshot (o riepilogo) del modello di cluster, compresa una misura della silhouette di coesione e separazione dei cluster, che è ombreggiata per indicare risultati scarsi, discreti o buoni. Questa snapshot consente di verificare rapidamente se la qualità è scarsa, nel qual caso è possibile decidere di tornare al nodo per la creazione dei modelli per correggere le impostazioni del modello di cluster e ottenere un risultato migliore.

La qualità del risultato (scarso, discreto, buono) è basata sul lavoro di Kaufman e Rousseeuw (1990) relativo all'interpretazione delle strutture dei cluster. Nella visualizzazione Riepilogo del modello, un risultato buono equivale a quei dati che rispecchiano la classificazione di Kaufman e Rousseeuw di ragionevole o forte indizio di una struttura di cluster, un risultato discreto rispecchia la classificazione di indizio debole, un risultato scarso corrisponde alla classificazione di assenza di indizio significativo.

La misura della silhouette calcola la media, su tutti i record, di (B–A) / max(A,B), dove A è la distanza del record dal centro del cluster relativo e B è la distanza del record dal centro del cluster più vicino a cui non appartiene. Un coefficiente di silhouette pari a 1 indica che tutti i casi si trovano direttamente in corrispondenza dei relativi centri di cluster. Il valore –1 indica che tutti i casi si trovano nei centri di altri cluster. Il valore 0 indica, in media, che i casi sono equidistanti tra il centro di cluster e il cluster più vicino.

Il riepilogo include una tabella che contiene le informazioni seguenti:

- **Algoritmo.** L'algoritmo di raggruppamento utilizzato (ad esempio, "TwoStep").
- **Funzioni di input.** Il numero di campi, noti anche come **input** o **predittori**.
- **Cluster.** Il numero di cluster nella soluzione.

Visualizzazione cluster

La visualizzazione Cluster contiene una griglia cluster-per-funzioni che comprende il nome, le dimensioni e il profilo di ciascun cluster.

Le colonne della griglia contengono le seguenti informazioni:

- **Cluster.** I numeri di cluster creati dall'algoritmo.
- **Etichetta.** L'eventuale etichetta applicata a ciascun cluster (che è vuota, per impostazione predefinita). Fare doppio clic nella cella per immettere un'etichetta che descrive il contenuto del cluster: ad esempio, "Acquirenti di auto di lusso".
- **Descrizione.** L'eventuale descrizione del contenuto del cluster (che è vuota, per impostazione predefinita). Fare doppio clic nella cella per immettere una descrizione del cluster: ad esempio, "età oltre i 55 anni, professionisti, reddito superiore a 100.000 euro".
- **Dimensione.** Le dimensioni di ciascun cluster sotto forma di percentuale dell'intero campione di cluster. Ogni cella relativa alle dimensioni all'interno della griglia visualizza una barra verticale che mostra la percentuale delle dimensioni all'interno del cluster, la percentuale delle dimensioni in formato numerico e il conteggio dei casi di cluster.
- **Caratteristiche.** I singoli input o predittori, ordinati per impostazione predefinita in base all'importanza globale. Se delle colonne hanno dimensioni uguali vengono mostrate in base al criterio di ordinamento crescente dei numeri di cluster.

  L'importanza generale di una funzione è indicata dal colore dell'ombreggiatura di sfondo della cella; la funzione più importante è la più scura, mentre quella meno importante è priva di ombreggiatura. Una guida al di sopra della tabella indica l'importanza associata al colore di ciascuna cella relativa a una funzione.

Quando si passa il mouse sopra una cella, vengono visualizzati il nome completo o l'etichetta della funzione e il valore di importanza della cella. È possibile che vengano visualizzate altre informazioni, a seconda della visualizzazione e del tipo di funzione. Nella visualizzazione Centri di cluster, include le statistiche ed il valore della cella; ad esempio: “Media: 4.32”. Per le funzioni relative alla categoria, la cella mostra il nome della categoria (modale) più frequente e la relativa percentuale.

All'interno della visualizzazione dei cluster, è possibile selezionare diversi metodi per visualizzare le informazioni sul cluster:

- Trasponi cluster e funzioni. Per ulteriori informazioni, consultare l'argomento “Trasponi cluster e funzioni”.
- Ordina funzioni. Per ulteriori informazioni, consultare l'argomento “Ordina funzioni”.
- Ordina cluster. Per ulteriori informazioni, consultare l'argomento “Ordina cluster”.
- Seleziona contenuto celle. Per ulteriori informazioni, consultare l'argomento “Contenuti cella”.

Trasponi cluster e funzioni: Per impostazione predefinita, i cluster vengono visualizzati sotto forma di colonne e le funzioni sotto forma di righe. Per invertire questa modalità, fare clic sul pulsante **Trasponi cluster e funzioni** a sinistra dei pulsanti **Ordina funzioni in base a**. Ad esempio, è possibile utilizzare questa opzione quando sono visualizzati troppi cluster, per ridurre la quantità di scorrimento orizzontale necessario per visionare i dati.

Ordina funzioni: I pulsanti **Ordina funzioni in base a** consentono di selezionare il modo in cui sono visualizzate le celle delle funzioni:

- **Importanza globale.** È l'impostazione predefinita. Le funzioni vengono organizzate in ordine di importanza globale decrescente, e il criterio di ordinamento è lo stesso tra i cluster. Se in qualche funzione sono presenti dei valori di importanza a pari merito, le funzioni a pari merito vengono elencate in base al criterio di ordinamento crescente in base ai nomi delle funzioni stesse.
- **Importanza entro i cluster.** Le funzioni vengono ordinate rispetto alla loro importanza per ciascun cluster. Se in qualche funzione sono presenti dei valori di importanza a pari merito, le funzioni a pari merito vengono elencate in base al criterio di ordinamento crescente in base ai nomi delle funzioni stesse. Quando si seleziona questa opzione, di solito il criterio di ordinamento varia tra i cluster.
- **Nome.** Le funzioni vengono ordinate alfabeticamente in base al nome.
- **Ordine dei dati.** Le funzioni vengono ordinate in base al loro ordine nell'insieme di dati.

Ordina cluster: Per impostazione predefinita, i cluster vengono ordinati in modo decrescente in base alla dimensione. I pulsanti **Ordina cluster in base a** consentono di ordinarli alfabeticamente per nome o, se sono state create delle etichette alfanumeriche univoche, rispetto a queste ultime.

Le funzioni con la stessa etichetta vengono ordinate in base al nome del cluster. Se i cluster sono ordinati in base alle etichette e si modifica l'etichetta di un cluster, il criterio di ordinamento viene aggiornato automaticamente.

Contenuti cella: I pulsanti **Celle** consentono di modificare la visualizzazione dei contenuti delle celle per quanto riguarda le funzioni e i campi di valutazione.

- **Centri di cluster.** Per impostazione predefinita, le celle visualizzano i nomi e le etichette delle funzioni e la tendenza centrale per ciascuna combinazione cluster/funzione. La media viene mostrata per i campi continui e la moda (categoria che ricorre più frequentemente) con la percentuale della categoria per i campi categoriali.
- **Distribuzioni assolute.** Mostra i nomi e le etichette e le distribuzioni assolute delle funzioni all'interno di ciascun cluster. Per le funzioni categoriali, la schermata visualizza dei grafici a barre a cui sono sovrapposte delle categorie ordinate in modo crescente rispetto ai valori dei dati. Per le funzioni continue, la schermata mostra un grafico di densità regolare che utilizza gli stessi punti finali e intervalli per ciascun cluster.

  La schermata in rosso pieno mostra la distribuzione dei cluster, mentre quella più chiara rappresenta i dati globali.
- **Distribuzioni relative.** Mostra i nomi e le etichette delle funzioni e le distribuzioni relative nelle celle. In generale, le schermate sono simili a quelle visualizzate per le distribuzioni assolute, a eccezione del fatto che vengono mostrate le distribuzioni relative.

  La schermata in rosso pieno mostra la distribuzione dei cluster, mentre quella più chiara rappresenta i dati globali.

- **Visualizzazione di base.** In presenza di molti cluster, può risultare difficile visualizzare i dettagli senza ricorrere allo scorrimento. Per ridurre la quantità di scorrimento, selezionare questa visualizzazione per passare a una versione più compatta della tabella.

Visualizzazione Importanza predittore nei cluster

La visualizzazione Importanza predittore mostra l'importanza relativa di ciascun campo nella stima del modello.

Visualizzazione Dimensioni dei cluster

La visualizzazione Dimensioni dei cluster mostra un grafico a torta che contiene ciascun cluster. La dimensione percentuale di ciascun cluster viene mostrata in ogni sezione; passare il mouse sopra ogni sezione per visualizzare il conteggio al suo interno.

Al di sotto del grafico, una tabella elenca le seguenti informazioni relative alle dimensioni:

- La dimensione del cluster più piccolo (sia il conteggio che una percentuale rispetto al totale).
- La dimensione del cluster più grande (sia il conteggio che una percentuale rispetto al totale).
- Il rapporto tra la dimensione del cluster più grande e quella del cluster più piccolo.

Visualizzazione Distribuzione delle celle

La visualizzazione Distribuzione delle celle mostra un grafico espanso e più dettagliato della distribuzione dei dati per qualsiasi cella di funzione selezionata nella tabella del riquadro principale dei cluster.

Visualizzazione Confronto tra cluster

La visualizzazione Confronto tra cluster è costituita da un layout a griglia, con le funzioni nelle righe e i cluster selezionati nelle colonne. Questa visualizzazione aiuta a comprendere meglio i fattori che formano i cluster; inoltre, consente di visualizzare le differenze tra i cluster non solo confrontandoli con i dati globali ma anche l'uno con l'altro.

Per selezionare i cluster da visualizzare, fare clic sulla parte superiore della colonna dei cluster nel riquadro principale Cluster. Fare clic tenendo premuto Ctrl o Maiusc per selezionare o deselezionare più di un cluster per il confronto.

Nota: è possibile selezionare fino a cinque cluster per la visualizzazione.

I cluster vengono mostrati nell'ordine in cui sono stati selezionati, mentre l'ordine dei campi è determinato dall'opzione **Ordina funzioni in base a**. Quando si seleziona **Importanza entro i cluster**, i campi vengono sempre ordinati in base all'importanza globale.

I grafici sullo sfondo mostrano le distribuzioni globali di ciascuna funzione:

- Le funzioni categoriali vengono visualizzate sotto forma di grafici a punti, dove la dimensione del punto indica la categoria più frequente/modale per ogni cluster (per funzione).
- Le funzioni continue vengono visualizzate sotto forma di grafici a scatole, che mostrano le mediane globali e gli intervalli interquartili.

Sovrapposti a queste visualizzazioni in secondo piano sono i grafici a scatole per i cluster selezionati:

- Nel caso delle funzioni continue, i marker punti e le linee orizzontali indicano la mediana e l'intervallo interquartile per ogni cluster.
- Ciascun cluster è rappresentato per mezzo di un colore diverso, mostrato nella parte superiore della visualizzazione.

Esplorazione del Visualizzatore cluster

Il Visualizzatore cluster è una schermata interattiva, È possibile:

- Selezionare un campo o un cluster per visualizzare ulteriori dettagli.

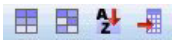
- Confrontare i cluster per selezionare gli elementi desiderati.
- Modificare la visualizzazione.
- Trasporre gli assi.
- Generare dei nodi Ricava, Filtro e Seleziona mediante il menu Genera.

Utilizzo delle barre degli strumenti

Le informazioni visualizzate nei riquadri destro e sinistro possono essere controllate mediante le opzioni delle barre degli strumenti. I controlli delle barre degli strumenti consentono di modificare l'orientamento della visualizzazione (dall'alto verso il basso, da sinistra verso destra o da destra verso sinistra). Inoltre, è anche possibile reimpostare le impostazioni predefinite del visualizzatore e aprire una finestra di dialogo per specificare il contenuto della visualizzazione Cluster nel riquadro principale.

Le opzioni **Ordina funzioni in base a**, **Ordina cluster in base a**, **Celle** e **Visualizza** sono disponibili solo quando si seleziona la visualizzazione **Cluster** nel riquadro principale. Per ulteriori informazioni, consultare l'argomento "Visualizzazione cluster" a pagina 223.

Tabella 12. Icone della barra degli strumenti.

Icona	Argomento
	Consultare Trasponi cluster e funzioni
	Consultare Ordina funzioni in base a
	Consultare ordina cluster in base a
	Consultare Celle

Generazione di nodi dai modelli di cluster

Il menu Genera consente di creare dei nuovi nodi in base al modello di cluster. Questa opzione è disponibile nella scheda Modello del modello generato e consente di generare dei nodi in base alla visualizzazione o alla selezione corrente (cioè, in base a tutti i cluster visibili o a tutti quelli selezionati). Ad esempio, è possibile selezionare una sola funzione e successivamente generare un nodo Filtro per scartare tutte le altre funzioni (non visibili). I nodi generati vengono posizionati senza collegamenti nell'area. Inoltre, è possibile generare una copia del nugget del modello nella palette Modelli. È importante ricordarsi di collegare i nodi e apportare le eventuali modifiche desiderate prima dell'esecuzione.

- **Genera nodo modellazione.** Crea un nodo per la creazione dei modelli sull'area del flusso. Questa opzione può rivelarsi utile, ad esempio, se si dispone di un flusso in cui si desidera utilizzare queste impostazioni di modello ma non si dispone più del nodo per la creazione dei modelli utilizzato per generarle.
- **Modello a palette.** Crea il nugget nella palette Modelli. Può rivelarsi utile in quelle situazioni in cui un collega abbia inviato un flusso che contiene il modello ma non il modello stesso.
- **Nodo filtro.** Crea un nuovo nodo Filtro per filtrare i campi che non sono utilizzati dal modello di cluster e/o non sono visibili nella schermata corrente del Visualizzatore cluster. Se è presente un nodo Tipo a monte per il nodo Cluster, tutti i campi con il ruolo *Obiettivo* vengono scartati dal nodo Filtro generato.
- **Nodo filtro (da selezione).** Crea un nuovo nodo Filtro per filtrare i campi in base alle selezioni effettuate nel Visualizzatore cluster. Per selezionare più campi, fare clic tenendo premuto Ctrl. I campi selezionati nel Visualizzatore cluster vengono scartati a valle, ma è possibile modificare questo comportamento modificando il nodo Filtro prima dell'esecuzione.

- **Nodo Seleziona.** Crea un nuovo nodo Seleziona per selezionare i record in base alla loro appartenenza a uno qualunque dei cluster visibili nella schermata del Visualizzatore cluster. Viene automaticamente generata una condizione di selezione.
- **Nodo Seleziona (da selezione).** Crea un nuovo nodo Seleziona per selezionare i record in base alla loro appartenenza a uno qualunque dei cluster selezionati nel Visualizzatore cluster. Per selezionare più cluster, fare clic tenendo premuto Ctrl.
- **Nodo Ricava.** Crea un nuovo nodo Ricava, che ottiene un campo flag che assegna ai record il valore *True* o *False* in base all'appartenenza a tutti i cluster visibili nel Visualizzatore cluster. Viene automaticamente generata una condizione di nuovo campo.
- **Nodo Ricava (da selezione).** Crea un nuovo nodo Ricava, che ottiene un campo flag in base all'appartenenza nei cluster selezionati nel Visualizzatore cluster. Per selezionare più cluster, fare clic tenendo premuto Ctrl.

Oltre che per generare i nodi, è possibile utilizzare il menu Genera anche per creare dei grafici. Per ulteriori informazioni, consultare l'argomento "Generazione di grafici dai modelli di cluster".

Controllo della visualizzazione cluster

Per controllare ciò che viene mostrato nella visualizzazione Cluster nel riquadro principale, fare clic sul pulsante **Visualizza**; viene aperta la finestra di dialogo Visualizza.

Caratteristiche. Selezionata per impostazione predefinita. Per nascondere tutte le funzioni di input, deselezionare la casella di controllo.

Campi di valutazione. Scegliere i campi di valutazione (campi non utilizzati per creare il modello di cluster, ma inviati al visualizzatore del modello per valutare i cluster) da visualizzare; per impostazione predefinita non ne è visualizzato nessuno. *Nota*: questa casella di spunta non è disponibile se non sono disponibili campi di valutazione.

Descrizioni cluster. Selezionata per impostazione predefinita. Per nascondere tutte le celle delle descrizioni dei cluster, deselezionare la casella di controllo.

Dimensioni dei cluster. Selezionata per impostazione predefinita. Per nascondere tutte le celle delle dimensioni dei cluster, deselezionare la casella di controllo.

Numero massimo di categorie. Specifica il numero massimo di categorie da visualizzare nei grafici delle funzioni categoriali; il valore predefinito è 20.

Generazione di grafici dai modelli di cluster

I modelli di cluster forniscono numerose informazioni; tuttavia, tali informazioni non sempre sono in un formato facilmente accessibile per gli utenti di business. Per fornire dati che siano facilmente incorporabili nei report di business, nelle presentazioni, e così via, è possibile produrre dei grafici dei dati selezionati. Ad esempio, dal Visualizzatore cluster è possibile generare un grafico per un cluster selezionato, creando in tal modo solo un grafico per i casi in tale cluster.

Nota: è possibile generare un grafico dal visualizzatore cluster solo quando il nugget del modello è collegato ad altri nodi in un flusso.

Generazione di un grafico

1. Aprire il nugget del modello che contiene il Visualizzatore cluster.
2. Nella scheda Modello, selezionare *Cluster* dall'elenco a discesa **Visualizza**.
3. Nella visualizzazione principale, selezionare uno o più cluster per cui si desidera produrre un grafico.
4. Dal menu Genera, selezionare **Grafico (da selezione)**; viene visualizzata la scheda Di base - Lavagna grafica.

Nota: quando si visualizza la Lavagna grafica in questo modo, sono disponibili soltanto le schede Di base e Dettagliato.

5. Mediante le impostazioni della scheda Di base o Dettagliato, specificare i dettagli da visualizzare sul grafico.
6. Fare clic su OK per generare il grafico.

L'intestazione del grafico identifica il tipo di modello e i cluster che sono stati scelti per essere inclusi.

Capitolo 12. Regole di associazione

Le **regole di associazione** consentono di associare una conclusione specifica (l'acquisto di un particolare prodotto) a un insieme di condizioni (l'acquisto di numerosi altri prodotti). Per esempio, la regola

```
birra <= verdura_scatola&carne_surgelata (173, 17.0%, 0.84)
```

indica che *birra* si riscontra spesso quando sono presenti contemporaneamente *verdura_scatola* e *carne_surgelata*. La regola è caratterizzata da un'affidabilità pari all'84% ed è applicabile al 17% dei dati, corrispondente a 173 record. Mediante gli algoritmi di regole di associazione vengono trovate automaticamente le associazioni che potrebbero essere individuate manualmente mediante tecniche di visualizzazione, come il nodo Web.

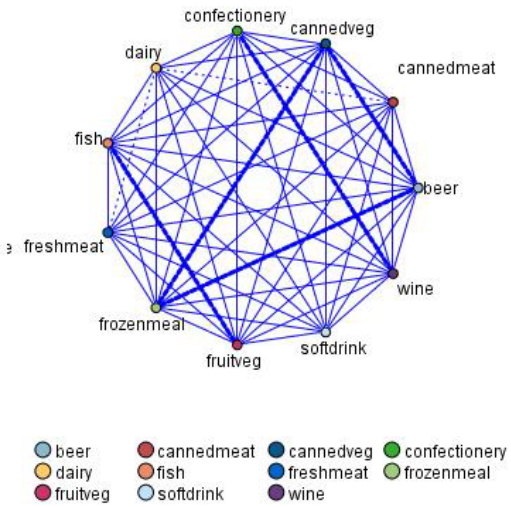


Figura 45. Nodo Web che mostra le associazioni tra gli elementi del market basket

Il vantaggio degli algoritmi delle regole di associazione rispetto agli algoritmi delle strutture ad albero delle decisioni più standard (C5.0 e C&R Trees) è che le associazioni possono esistere tra *qualsiasi* attributo. Mentre un algoritmo della struttura ad albero delle decisioni genera regole con un'unica conclusione, gli algoritmi di associazione tentano di individuare più regole, ciascuna delle quali può fornire una diversa conclusione.

Gli algoritmi di associazione presentano tuttavia lo svantaggio di tentare di individuare schemi in un ambito di ricerca potenzialmente molto esteso e quindi di essere caratterizzati da tempi di esecuzione notevolmente superiori rispetto a un algoritmo della struttura ad albero delle decisioni. Per individuare le regole, gli algoritmi utilizzano un metodo di **generazione e test** con cui vengono inizialmente generate regole semplici, che vengono quindi convalidate rispetto all'insieme di dati. Le regole corrette vengono archiviate e tutte le regole, soggette a diversi vincoli, vengono quindi sottoposte a specializzazione. **specializzazione** è il processo con cui vengono aggiunte condizioni a una regola. Queste nuove regole vengono quindi convalidate rispetto ai dati e il processo memorizza in modo iterativo le regole più efficienti o interessanti individuate. Per il numero di antecedenti ammesso in una regola viene in genere specificato dall'utente un limite massimo, mentre per ridurre l'ambito di ricerca potenzialmente esteso vengono utilizzate tecniche basate sulla teoria dell'informazione e su schemi di indicizzazione efficienti.

Al termine dell'elaborazione viene presentata una tabella delle regole migliori. A differenza di una struttura ad albero delle decisioni, questo insieme di regole di associazione non può essere direttamente utilizzato per elaborare previsioni come un modello standard, quale una struttura ad albero delle decisioni o una rete neurale, poiché per le regole possono essere ottenute più conclusioni diverse. Per

trasformare le regole di associazione in un insieme di regole di classificazione è richiesto un ulteriore livello di trasformazione. Per questo motivo, le regole di associazione generate dagli algoritmi di associazione vengono definite **modelli grezzi**. Sebbene possano essere esplorati dall'utente, i modelli grezzi non possono essere utilizzati esplicitamente come modelli di classificazione, a meno che l'utente non richieda al sistema di generare un modello di classificazione dal modello grezzo. Questa operazione può essere eseguita mediante la voce di menu Genera disponibile nel browser.

Sono supportati due algoritmi di regole di associazione:

Il nodo Apriori estrae un insieme di regole dai dati, estrapolando le regole con il più alto contenuto di informazioni. Apriori offre cinque diversi metodi per la selezione delle regole e utilizza uno schema di indicizzazione sofisticato per elaborare in modo efficiente insiemi di dati di grandi dimensioni. In caso di problemi complessi, l'addestramento di Apriori è in genere più rapido. Apriori non ha un limite arbitrario per quanto riguarda il numero di regole che possono essere mantenute e può gestire regole con un massimo di 32 precondizioni. Apriori richiede che tutti i campi di input e output siano categoriali ma garantisce prestazioni migliori perché è ottimizzato per questo tipo di dati.

Il nodo Sequenza consente di scoprire le regole di associazione nei dati sequenziali o basati su valori temporali. Per sequenza si intende un elenco di serie di elementi che tendono a ricorrere secondo un ordine prevedibile. Ad esempio, un cliente che acquista un rasoio e la lozione dopobarba potrebbe in seguito acquistare la schiuma da barba. Il nodo Sequenza si basa sull'algoritmo delle regole di associazione CARMA, che utilizza un metodo efficiente in due passaggi per trovare le sequenze.

Dati in formato tabellare e dati transazionali

I dati utilizzati dai modelli di regole di associazione possono essere in formato transazionale o tabulare, come illustrato di seguito. Di seguito vengono riportate descrizioni generali. Per requisiti specifici, si rimanda alla documentazione specifica relativa a ogni tipo di modello. Si noti che durante il calcolo del punteggio di modelli, i dati di cui si calcola il punteggio devono rispecchiare il formato dei dati utilizzati per creare il modello. È possibile utilizzare i modelli generati utilizzando dati in formato tabellare solo per calcolare il punteggio di dati in formato tabellare. Viceversa, i modelli generati utilizzando dati transazionali possono essere utilizzati solo per calcolare il punteggio di dati transazionali.

Formato transazionale

I dati transazionali presentano un record separato per ogni transazione o elemento. Se un cliente effettua più acquisti, per esempio, ogni acquisto costituirà un record separato, con elementi associati collegati da un ID cliente. Questo formato è conosciuto anche come formato di **registro incrementale acquisti**.

Cliente	Acquisto
1	marmellata
2	latte
3	marmellata
3	pane
4	marmellata
4	pane
4	latte

I nodi Apriori, CARMA e Sequenza possono utilizzare tutti dati transazionali.

Dati in formato tabellare

I dati in formato tabellare (noti anche come dati **basket** o di **tabella verità**) dispongono di elementi rappresentati da flag separati, dove ogni campo flag rappresenta la presenza o l'assenza di un determinato elemento. Ogni record rappresenta un insieme completo di elementi associati. I campi flag possono essere categoriali o numerici, sebbene determinati modelli potrebbero avere requisiti più specifici.

Cliente	Marmellata	Pane	Latte
1	T	F	F
2	F	F	T
3	T	T	F
4	T	T	T

I nodi Apriori, CARMA, e i nodi Sequenza possono utilizzare i dati in formato tabellare.

Nodo Apriori

Il nodo Apriori scopre anche le regole di associazione presenti nei dati. Apriori offre cinque diversi metodi per la selezione delle regole e utilizza uno schema di indicizzazione sofisticato per elaborare in modo efficiente insiemi di dati di grandi dimensioni.

Requisiti. Per creare un insieme di regole Apriori, è necessario avere uno o più campi *Input* e uno o più campi *Obiettivo*. I campi di input e di output (quelli con il ruolo *Input*, *Obiettivo* o *Entrambi*) devono essere simbolici. I campi con il ruolo *Nessuna* verranno ignorati. I tipi dei campi devono essere completamente istanziati prima dell'esecuzione del nodo. I dati possono essere nel formato tabulare o transazionale. Per ulteriori informazioni, consultare l'argomento “Dati in formato tabellare e dati transazionali” a pagina 230.

Efficacia. In caso di problemi complessi, l'addestramento di Apriori in genere è più rapido. Inoltre non ha un limite arbitrario per quanto riguarda il numero di regole che possono essere mantenute e può gestire regole con un massimo di 32 precondizioni. Apriori offre cinque diversi metodi di addestramento, permettendo una maggiore flessibilità nell'associazione del metodo di data mining al problema da risolvere.

Opzioni del modello di nodo Apriori

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.

Supporto minimo regola. È possibile specificare un criterio di supporto per mantenere le regole nell'insieme di regole. Per **supporto** si intende la percentuale dei record nei dati di addestramento i cui antecedenti (ovvero la parte della regola introdotta da "se") sono veri. Si noti che questa definizione del termine supporto è diversa da quella utilizzata nei nodi CARMA e Sequenza. Per ulteriori informazioni, consultare l'argomento “Opzioni del modello di nodo Sequenza” a pagina 246. Se si ottengono regole che possono essere applicate a un numero di sottoinsiemi di dati molto ridotto, tentare di aumentare il valore di questa impostazione.

Nota: la definizione del supporto per Apriori è basata sul numero di record con gli antecedenti. Ciò è in contrasto con gli algoritmi CARMA e Sequenza, per i quali tale definizione si basa sul numero dei record con tutti gli elementi di una regola, ovvero sia gli antecedenti che il conseguente. I risultati dei modelli di associazione mostrano le misure sia del supporto (antecedente) sia del supporto della regola.

Confidenza minima regola. È anche possibile specificare un criterio di confidenza. La **confidenza** si basa sui record per i quali gli antecedenti della regola sono veri ed è la percentuale dei record per i quali

anche i conseguenti sono veri. In altre parole, è la percentuale di previsioni corrette basate sulla regola. Le regole con confidenza inferiore al criterio specificato verranno scartate. Se si ottengono troppe regole, provare ad aumentare il valore specificato per questa impostazione. Se si ottengono poche o nessuna regola, provare a diminuire il valore di questa impostazione.

Numero massimo di antecedenti. È possibile specificare il numero massimo di precondizioni per qualsiasi regola. In questo modo è possibile ridurre la complessità delle regole. Se le regole sono troppo complesse o troppo specifiche, provare a diminuire il valore di questa impostazione. Questa impostazione influisce sui tempi di addestramento. Se per l'insieme di regole creato è necessario un tempo di addestramento eccessivo, provare a diminuire il valore di questa impostazione.

Solo valori veri per i flag. Se l'opzione è selezionata per i dati in formato tabulare (tabella di verità), nelle regole risultanti verranno inclusi unicamente i valori veri. Di conseguenza, potrebbe essere più facile capire le regole. L'opzione non si applica ai dati in formato transazionale. Per ulteriori informazioni, consultare l'argomento "Dati in formato tabellare e dati transazionali" a pagina 230.

Ottimizza. Selezionare le opzioni per l'incremento delle prestazioni durante la creazione del modello in base alle specifiche esigenze.

- Selezionare **Velocità** per indicare all'algoritmo di non utilizzare mai il riversamento su disco per migliorare le prestazioni.
- Selezionare **Memoria** per indicare all'algoritmo di utilizzare il riversamento su disco quando necessario anche se ciò può comportare una diminuzione delle prestazioni. Questa opzione è selezionata per default. *Nota*: durante l'esecuzione in modalità distribuita, questa impostazione può essere sovrascritta dalle opzioni dell'amministratore specificate in *options.cfg*. Per ulteriori informazioni, vedere il documento *IBM SPSS Modeler Server Administrator's Guide*.

Opzioni avanzate del nodo Apriori

Le seguenti opzioni avanzate consentono agli esperti del nodo Apriori di ottimizzare il processo di induzione. Per accedere alle opzioni avanzate, impostare la modalità su **Livello avanzato** nella scheda Livello avanzato.

Misura di valutazione. Apriori supporta cinque metodi di valutazione delle regole potenziali.

- **Confidenza regola.** Il metodo di default utilizza la confidenza (o precisione) della regola per valutare le regole. Per questa misura, **Limite inferiore della misura di valutazione** è disattivato, poiché è ridondante con l'opzione **Confidenza minima regola** della scheda Modello. Per ulteriori informazioni, consultare l'argomento "Opzioni del modello di nodo Apriori" a pagina 231.
- **Differenza confidenza.** (chiamata anche **differenza assoluta di confidenza rispetto a probabilità a priori**). Questa misura di valutazione è la differenza assoluta tra la confidenza della regola e la sua confidenza a priori. Questa opzione evita distorsioni quando i risultati non sono distribuiti in modo uniforme. Ciò consente di evitare che vengano conservate regole "ovvie". Ad esempio, potrebbe verificarsi il caso in cui l'80% dei clienti compri il prodotto più popolare di un'azienda. Una regola che prevede l'acquisto di quel prodotto con un'accuratezza dell'85% non aggiunge molto alle proprie conoscenze, anche se un'accuratezza dell'85% potrebbe sembrare ottima in valore assoluto. Impostare il limite inferiore della misura di valutazione sulla differenza minima nella confidenza per la quale si desidera mantenere le regole.
- **Rapporto confidenza.** (chiamato anche **differenza del quoziente di confidenza a 1**). Questa misura di valutazione è il rapporto tra la regola di confidenza e la confidenza a priori (oppure, se il rapporto è maggiore di uno, il suo reciproco) sottratto da 1. Analogamente alla Differenza confidenza, questo metodo prende in considerazione le distribuzioni irregolari. È particolarmente adatto per individuare le regole che prevedono eventi rari. Si supponga, ad esempio, che vi sia una rara condizione medica che colpisce solo l'1% dei pazienti. Una regola in grado di prevedere questa condizione nel 10% dei casi rappresenta un grande miglioramento rispetto ad ipotesi casuali, anche se, in una scala assoluta, un'accuratezza del 10% potrebbe non sembrare un ottimo risultato. Impostare il limite inferiore della misura di valutazione sulla differenza per la quale si vuole che le regole vengano mantenute.

- **Differenza informazioni.** (chiamata anche **differenza di informazioni rispetto a probabilità a priori**). Questa misura si basa sulla misura **guadagno di informazioni**. Se la probabilità di un particolare conseguente viene considerata un valore logico (un **bit**), il guadagno di informazioni è la percentuale di quel bit che può essere determinata, basata sugli antecedenti. La differenza di informazioni è la differenza tra il guadagno di informazioni, dati gli antecedenti, e il guadagno di informazioni, data solo la confidenza a priori del conseguente. Una particolarità importante di questo metodo è la sua valutazione del supporto, per cui le regole che coprono più record vengono preferite per un dato livello di confidenza. Impostare il limite inferiore della misura di valutazione sulla differenza di informazioni per cui si desidera che le regole siano mantenute.

  Nota: poiché la scala per questa misura è in qualche modo meno intuitiva rispetto ad altre scale, potrebbe essere necessario sperimentare limiti inferiori differenti per ottenere un insieme di regole soddisfacente.
- **Chi-quadrato normalizzato.** (chiamata anche **misura chi-quadrato normalizzato**). Questa misura è un indice statistico di associazioni tra antecedenti e conseguenti. La misura viene normalizzata per utilizzare i valori tra 0 e 1 e dipende dal supporto molto più della misura differenza di informazioni. Impostare il limite inferiore della misura di valutazione sulla differenza di informazioni per cui si desidera che le regole siano mantenute.

  Nota: come per la misura della differenza delle informazioni, la scala per questa misura è in qualche modo meno intuitiva rispetto alla altre scale, per cui potrebbe essere necessario sperimentare limiti inferiori differenti per ottenere un insieme di regole soddisfacente.

Consenti regole senza antecedenti. Selezionare questa opzione per consentire le regole che includono unicamente il conseguente (elemento o serie di elementi). È utile se si desidera definire elementi o serie di elementi comuni. Per esempio, `verdura_scatola` è una regola a elemento singolo senza un antecedente che indica che l'acquisto di *verdura_scatola* è un'occorrenza comune nei dati. In alcuni casi, è possibile inserire regole di questo tipo se si è interessati unicamente alle previsioni con la maggiore confidenza. Questa opzione è disattivata per default. Per convenzione, il supporto antecedente per le regole senza antecedenti è rappresentato come 100% e il supporto della regola sarà uguale alla confidenza.

Nodo CARMA

Il nodo CARMA utilizza un algoritmo di rilevamento delle regole di associazione per individuare le regole di associazione presenti nei dati. Le regole di associazione sono istruzioni nella forma

if *antecedent(s)* **then**
consequent(s)

Per esempio, se un cliente Web acquista una scheda senza fili e un router senza fili ad alte prestazioni, è probabile che acquisti anche un server musicale senza fili, se gli viene offerto. Il modello CARMA estrae un insieme di regole dai dati senza che venga richiesto all'utente di specificare i campi di input o obiettivo. Pertanto, le regole generate possono essere utilizzate per una gamma più vasta di applicazioni. Per esempio, mediante le regole generate dal nodo è possibile trovare un elenco di prodotti o di servizi (antecedenti) il cui conseguente è rappresentato dall'articolo che si desidera promuovere per le festività correnti. In IBM SPSS Modeler è possibile individuare i clienti che hanno acquistato i prodotti antecedenti e realizzare una campagna marketing destinata alla promozione del prodotto conseguente.

Requisiti. A differenza di Apriori, il nodo CARMA non richiede campi *Input* o *Obiettivo*. Ciò dipende dal tipo di funzionamento dell'algoritmo ed equivale a creare un modello Apriori con tutti i campi impostati su *Entrambi*. È possibile vincolare gli elementi che vengono elencati unicamente come antecedenti o conseguenti, filtrando il modello dopo che è stato creato. Per esempio, mediante il browser di modelli è possibile trovare un elenco di prodotti o di servizi (antecedenti) il cui conseguente è rappresentato dall'articolo che si desidera promuovere per le festività correnti.

Per creare un insieme di regole CARMA è necessario specificare un campo ID e uno o più campi contenuto. Il campo ID può avere qualsiasi ruolo o livello di misurazione. I campi con il ruolo *Nessuna* verranno ignorati. I tipi dei campi devono essere completamente istanziati prima dell'esecuzione del

nodo. Come per Apriori, il formato dei dati può essere tabulare o transazionale. Per ulteriori informazioni, consultare l'argomento “Dati in formato tabellare e dati transazionali” a pagina 230.

Efficacia. Il nodo CARMA si basa sull'algoritmo delle regole di associazione CARMA. A differenza di Apriori, il nodo CARMA fornisce le impostazioni di creazione per il supporto regola (sia per l'antecedente che per il conseguente) anziché il supporto antecedente. Il nodo CARMA consente inoltre le regole con più conseguenti. Analogamente ad Apriori, i modelli generati da un nodo CARMA possono essere inseriti in un flusso di dati allo scopo di creare previsioni. Per ulteriori informazioni, consultare l'argomento “Nugget del modello” a pagina 37.

Opzioni dei campi del nodo CARMA

Prima di eseguire un nodo CARMA è necessario specificare i campi di input nella scheda Campi del nodo CARMA. Mentre la maggior parte dei nodi Modelli condivide le stesse opzioni della scheda Campi, molte delle opzioni del nodo CARMA sono univoche. Tutte le opzioni verranno illustrate di seguito

Utilizza impostazioni nodo tipologia. Questa opzione indica al nodo di utilizzare le informazioni sui campi da un nodo tipologia 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 tipologia a monte. Dopo avere selezionato l'opzione, specificare i campi seguenti a seconda che si leggano i dati in formato transazionale o tabulare.

Utilizza formato transazionale. Questa opzione modifica gli altri campi disponibili nella finestra di dialogo a seconda che i dati siano in formato transazionale o tabulare. Se si utilizzano più campi multipli con dati transazionali, si presume che gli elementi specificati in questi campi per un particolare record rappresentino gli elementi trovati in una singola transazione con un singolo timestamp. Per ulteriori informazioni, consultare l'argomento “Dati in formato tabellare e dati transazionali” a pagina 230.

Dati in formato tabellare

Se non è selezionato **Utilizza formato transazionale**, vengono visualizzati i campi seguenti.

- **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.
- **Partizione.** Questo campo consente di specificare un campo utilizzato per partizionare i dati in campioni distinti per le fasi di addestramento, di test e di convalida della creazione del modello. Utilizzando un campione per generare il modello e un altro campione per sottoporlo a verifica, è possibile ottenere un'indicazione valida del modo in cui il modello potrà essere esteso a insiemi di dati di dimensioni maggiori, simili ai dati correnti. Se sono stati definiti più campi partizione utilizzando nodi Tipo o Partizione, 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. Si noti inoltre che per applicare la partizione selezionata nell'analisi, è necessario che il partizionamento sia attivato anche nella scheda Opzioni modello relativa al nodo. La disattivazione di questa opzione consente di disattivare il partizionamento senza dover modificare le impostazioni dei campi.

Dati transazionali

Se si seleziona **Utilizza formato transazionale**, vengono visualizzati i campi seguenti.

- **ID.** Per i dati transazionali, selezionare un campo ID dall'elenco. Come campo ID è possibile utilizzare campi numerici o simbolici. Ogni valore univoco di questo campo deve indicare una specifica unità di analisi. In un'applicazione per analisi market basket, per esempio, ciascun ID potrebbe rappresentare un singolo cliente. Per un'applicazione per analisi di registri Web, ciascun ID potrebbe rappresentare un computer (per indirizzo IP) o un utente (per dati di login).

- **ID contigui.** (Solo nodi Apriori e CARMA) Se i dati sono preordinati in modo che tutti i record con lo stesso ID vengano raggruppati 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.

  Nota: se i dati non sono ordinati e viene selezionata questa opzione, potrebbero essere restituiti risultati non validi nel modello.
- **Contenuto.** Specificare il campo o i campi contenuto per il modello. Questi campi contengono gli elementi rilevanti nella creazione di modelli di associazione. È possibile specificare più campi flag se i dati sono in formato tabulare o un singolo campo nominale se i dati sono in formato transazionale.

Opzioni del modello di nodo CARMA

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.

Supporto minimo regola (%). È possibile specificare un criterio di supporto. Per **supporto regola** si intende la proporzione di ID nei dati di addestramento che contengono l'intera regola. Si noti che questa definizione del termine supporto è diversa dal supporto antecedente utilizzato nei nodi Apriori. Per prendere in esame le regole più comuni, aumentare il valore di questa impostazione.

Confidenza minima regola (%). È possibile specificare un criterio di confidenza per mantenere le regole nell'insieme di regole. Per **confidenza** si intende la percentuale di ID in cui viene eseguita una previsione corretta, tra tutti gli ID per i quali la regola fa una previsione. La confidenza viene calcolata come numero di ID per i quali viene trovata l'intera regola, diviso per il numero di ID per i quali si trovano gli antecedenti, in base ai dati di addestramento. Le regole con confidenza inferiore al criterio specificato verranno scartate. Se si ottengono troppe regole o regole non di interesse, provare ad aumentare il valore specificato per questa impostazione. Se le regole ottenute sono troppo poche, provare a diminuire il valore di questa impostazione.

Dimensione massima regola. È possibile impostare il numero massimo di *serie di elementi* distinte (in contrapposizione agli *elementi*) in una regola. Se le regole significative sono relativamente brevi, è possibile diminuire il valore di questa impostazione per accelerare la creazione dell'insieme di regole.

Opzioni avanzate del nodo CARMA

Le opzioni avanzate riportate di seguito consentono agli esperti del nodo CARMA di ottimizzare il processo di generazione dei modelli. Per accedere alle opzioni avanzate, impostare la modalità su **Livello avanzato** nella scheda Livello avanzato.

Escludi regole con più conseguenti. Selezionare per escludere i "conseguenti con due elementi". Per esempio, la regola `pane & formaggio & pesce -> vino&frutta` contiene un conseguente con due elementi, `vino&frutta`. Per default, queste regole vengono incluse.

Imposta valore di taglio. Durante l'elaborazione, l'algoritmo CARMA utilizzato rimuove periodicamente (**taglia**) le serie di elementi non frequenti dall'elenco delle serie di elementi potenziali per preservare la memoria. Selezionare questa opzione per modificare la frequenza delle operazioni di taglio. Il numero specificato determina la frequenza di taglio. Specificare un valore minore per diminuire i requisiti di memoria dell'algoritmo (aumentando tuttavia il tempo di addestramento richiesto), oppure specificare un valore maggiore per accelerare l'addestramento (tenendo tuttavia presente che questa impostazione può aumentare i requisiti di memoria). Il valore predefinito è 500.

Varia supporto. Selezionare questa opzione per aumentare l'efficienza escludendo le serie di elementi non frequenti che risultano come frequenti perché sono incluse in modo irregolare. A tale scopo, partire da un

livello di supporto più alto e abbassarlo fino al livello specificato nella scheda Modello. Nel campo **Numero di transazioni stimato** specificare un valore che indica la rapidità con la quale deve essere abbassato il livello di supporto.

Consenti regole senza antecedenti. Selezionare questa opzione per consentire le regole che includono unicamente il conseguente (elemento o serie di elementi). È utile se si desidera definire elementi o serie di elementi comuni. Per esempio, `verdura_scatola` è una regola a elemento singolo senza un antecedente che indica che l'acquisto di *verdura_scatola* è un'occorrenza comune nei dati. In alcuni casi, è possibile inserire regole di questo tipo se si è interessati unicamente alle previsioni con la maggiore confidenza. Questa opzione è deselezionata per default.

Nugget del modello di regole di associazione

I nugget del modello di regole di associazione rappresentano le regole individuate da uno dei seguenti nodi Modelli delle regole di associazione:

- Apriori
- CARMA

I nugget del modello contengono le informazioni relative alle regole estratte dai dati durante la creazione del modello.

Visualizzazione di risultati

È possibile sfogliare le regole generate dai modelli di associazione (Apriori e CARMA) e i modelli Sequenza utilizzando la scheda Modello disponibile nella finestra di dialogo. Sfogliando un nugget del modello è possibile visualizzare le informazioni sulle regole e le opzioni di filtro e di ordinamento dei risultati prima di passare alla generazione di nuovi nodi o al calcolo del punteggio del modello.

Calcolo del punteggio del modello

È possibile aggiungere al flusso nugget del modello rifiniti (Apriori, CARMA e Sequenza) e utilizzarli per il calcolo del punteggio. Per ulteriori informazioni, consultare l'argomento "Utilizzo dei nugget del modello nei flussi" a pagina 48. I nugget del modello utilizzati per il calcolo del punteggio includono una scheda Impostazioni supplementare nelle rispettive finestre di dialogo. Per ulteriori informazioni, consultare l'argomento "Impostazioni del nugget del modello di regole di associazione" a pagina 240.

Un nugget del modello grezzo non può essere utilizzato per il calcolo del punteggio. In alternativa, è possibile generare un insieme di regole e utilizzarlo per il calcolo del punteggio. Per ulteriori informazioni, consultare l'argomento "Generazione di un insieme di regole da un nugget del modello di associazione" a pagina 241.

Dettagli del nugget del modello di regole di associazione

Nella scheda Modello di un nugget del modello di regole di associazione, è possibile visualizzare una tabella contenente le regole estratte dall'algoritmo. Ogni riga nella tabella rappresenta una regola. La prima colonna rappresenta i conseguenti (la parte "then" della regola) e la colonna successiva rappresenta gli antecedenti (la parte "if" della regola). Le colonne successive contengono informazioni sulla regola, per esempio la confidenza, il supporto e il guadagno cumulativo.

Le regole di associazione sono spesso visualizzate nel formato indicato nella tabella riportata di seguito.

Tabella 13. Esempio di regola di associazione

Conseguente	Antecedente
`Drug = drugY`	`Sex = F` `BP = HIGH`

La regola di esempio viene interpretata nel modo seguente *se Sesso= "F" e Pressione = "ALTO", è probabile che Cura sia curaY* oppure, in altre parole, *per i record dove Sesso = "F" e Pressione = "ALTO", è probabile che Cura sia curaY*. Mediante la barra degli strumenti disponibile nella finestra di dialogo, è possibile scegliere di visualizzare informazioni aggiuntive, per esempio la confidenza, il supporto e le istanze.

Menu Ordina. Il pulsante del menu Ordina sulla barra degli strumenti controlla l'ordinamento delle regole. La direzione dell'ordinamento (crescente o decrescente) può essere modificata utilizzando il pulsante di direzione dell'ordinamento (freccia in su o in giù).

È possibile ordinare le regole in base a:

- Supporto
- Confidenza
- Supporto regola
- Conseguente
- Guadagno cumulativo
- Distribuibilità

Menu Mostra/Nascondi. Il menu Mostra/Nascondi (pulsante della barra degli strumenti dei criteri) controlla le opzioni di visualizzazione delle regole.

Figura 46. Pulsante Mostra/Nascondi

Sono disponibili le seguenti opzioni di visualizzazione:

- **ID regola** visualizza l'ID regola assegnato durante la creazione del modello. L'ID regola consente di identificare le regole applicate per una previsione specifica e inoltre di unire successivamente informazioni aggiuntive sulle regole, per esempio la capacità di distribuzione, le informazioni sul prodotto o gli antecedenti.
- **Istanze** visualizza informazioni sul numero di ID univoci ai quali è applicata la regola, ovvero per i quali gli antecedenti sono veri. Per esempio, data la regola `pane -> formaggio`, il numero di record dei dati di addestramento che includono l'antecedente *pane* è definito come **istanze**.
- **Supporto** visualizza il supporto antecedente — ovvero la proporzione di ID per cui gli antecedenti sono veri, in base ai dati di addestramento. Per esempio, se il 50% dei dati di addestramento include l'acquisto di pane, allora la regola `pane -> formaggio` avrà un supporto antecedente pari al 50%. *Nota*: il supporto definito in questo punto è uguale alle istanze, ma è rappresentato come percentuale.
- **Confidenza** visualizza il rapporto tra il supporto regola e il supporto antecedente. Questo rapporto indica la proporzione di ID con gli antecedenti specificati per i quali anche i conseguenti sono veri. Per esempio, se il 50% dei dati di addestramento contiene il pane (a indicare il supporto antecedente) ma solo il 20% contiene sia il pane che il formaggio (a indicare il supporto della regola), allora la confidenza per la regola `pane -> formaggio` sarà `Supporto regola/Supporto antecedente` oppure, in questo caso, il 40%.
- **Supporto regola** visualizza la proporzione di ID per i quali l'intera regola, gli antecedenti e i conseguenti sono veri. Ad esempio, se il 20% dei dati di addestramento contiene pane e formaggio, il supporto della regola per la regola `pane -> formaggio` è 20%.
- **Guadagno cumulativo** visualizza il rapporto tra la confidenza per la regola e la probabilità a priori di avere il conseguente. Per esempio, se il 10% dell'intera popolazione acquista pane, allora una regola che prevede se le persone acquisteranno il pane con una confidenza del 20% avrà un guadagno cumulativo pari a 20/10 = 2. Se un'altra regola indica che le persone acquisteranno il pane con una confidenza dell'11%, allora la regola avrà un guadagno cumulativo vicino a 1, a indicare che la

presenza degli antecedenti non influisce molto sulla probabilità di avere il conseguente. In generale, le regole con un guadagno cumulativo diverso da 1 saranno più interessanti delle regole con un guadagno cumulativo vicino a 1.

- **Distribuibilità** è una misura della percentuale dei dati di addestramento che soddisfa le condizioni dell'antecedente, ma che non soddisfa il conseguente. In termini di prodotti acquistati, indica la percentuale della base di clienti totale che ha (o ha acquistato) gli antecedenti ma non ha ancora acquistato il conseguente. La statistica di distribuibilità è definita come `((Supporto antecedente in n. di record - Supporto regola in n. di record) / Numero di record) * 100`, dove *Supporto antecedente* rappresenta il numero di record per i quali sono veri gli antecedenti e *Supporto regola* rappresenta il numero di record per i quali sono veri sia gli antecedenti sia il conseguente.

Pulsante Filtro. Il pulsante Filtro (icona a forma di imbuto) sul menu espande la parte inferiore della finestra di dialogo e visualizza un riquadro che contiene i filtri della regola attivi. I filtri consentono di limitare il numero delle regole visualizzate nella scheda Modelli.

Figura 47. Pulsante Filtro

Per creare un filtro, fare clic sull'icona Filtro a destra del riquadro espanso. Verrà aperta una finestra di dialogo in cui è possibile specificare i vincoli per la visualizzazione delle regole. Si noti che il pulsante Filtro viene spesso utilizzato insieme al menu Genera per filtrare innanzitutto le regole e creare quindi un modello che contiene tale sottoinsieme di regole. Per ulteriori informazioni, consultare “Definizione dei filtri per le regole” di seguito.

Pulsante Trova regola. Il pulsante Trova regola (icona a forma di binocolo) consente di cercare le regole visualizzate per l'ID regola specificato. Nella casella adiacente è indicato il numero delle regole correntemente visualizzate. Gli ID regola vengono assegnati dal modello in base all'ordine di rilevamento corrente e vengono aggiunti ai dati durante il calcolo del punteggio.

Figura 48. Pulsante Trova regola

Per riordinare gli ID delle regole:

1. Per riorganizzare gli ID delle regole in IBM SPSS Modeler è possibile innanzitutto ordinare la tabella di visualizzazione delle regole in base alla misura desiderata, per esempio la confidenza o il guadagno cumulativo.
2. Utilizzare quindi le opzioni del menu Genera per creare un modello filtrato.
3. Nella finestra di dialogo Modello filtrato, selezionare **Rinumera regole consecutivamente a partire da** e specificare un numero.

Per ulteriori informazioni, consultare l'argomento “Generazione di un modello filtrato” a pagina 241.

Definizione dei filtri per le regole

Per default, gli algoritmi di regole quali Apriori, CARMA e Sequenza possono generare un numero eccessivo di regole. Per semplificare il processo di esplorazione o il calcolo del punteggio delle regole, è consigliabile filtrare le regole in modo da visualizzare in modo appropriato i conseguenti e gli antecedenti desiderati. Mediante le opzioni di filtro disponibili nella scheda Modello di un browser di regole, è possibile aprire una finestra di dialogo per la definizione delle impostazioni del filtro.

Conseguenti. Selezionare **Attiva filtro** per attivare le opzioni di filtro delle regole in base all'inclusione o all'esclusione dei conseguenti specificati. Selezionare **Includi uno** per creare un filtro in base al quale le regole contengono almeno uno dei conseguenti specificati. In alternativa, selezionare **Escludi** per creare

un filtro che esclude i conseguenti specificati. Per selezionare i conseguenti, è possibile utilizzare il pulsante di selezione a destra della casella di riepilogo, che consente di aprire una finestra di dialogo in cui sono elencati tutti i conseguenti presenti nelle regole generate.

Note: i conseguenti possono contenere più di un elemento. I filtri verificheranno unicamente che un conseguente contenga uno degli elementi specificati.

Antecedenti. Selezionare **Attiva filtro** per attivare le opzioni di filtro delle regole in base all'inclusione o all'esclusione degli antecedenti specificati. Per selezionare gli elementi, è possibile utilizzare il pulsante di selezione a destra della casella di riepilogo, che consente di aprire una finestra di dialogo in cui sono elencati tutti gli antecedenti presenti nelle regole generate.

- Selezionare **Includi tutti** per definire un filtro in base al quale tutti gli antecedenti specificati verranno inclusi in una regola.
- Selezionare **Includi uno** per creare un filtro in base al quale le regole contengono almeno uno dei antecedenti specificati.
- Selezionare **Escludi** per creare un filtro che esclude le regole contenenti un antecedente specificato.

Confidenza. Selezionare **Attiva filtro** per attivare le opzioni di filtro delle regole in base al livello di confidenza di una regola. Per definire un intervallo di confidenza è possibile utilizzare i controlli **Min** e **Max**. Quando si sfogliano i modelli generati, la confidenza è visualizzata come percentuale. Quando si calcola il punteggio dell'output, la confidenza è espressa da un numero compreso tra 0 e 1.

Supporto antecedente. Selezionare **Attiva filtro** per attivare le opzioni di filtro delle regole in base al livello di supporto antecedente di una regola. Il supporto antecedente indica la proporzione dei dati di addestramento che contiene gli stessi antecedenti della regola corrente, rendendola analoga a un indice di popolarità. È possibile definire un intervallo per il filtro delle regole in base al livello di supporto utilizzando i controlli **Min** e **Max**.

Guadagno cumulativo. Selezionare **Attiva filtro** per attivare le opzioni di filtro delle regole in base alla misurazione del guadagno cumulativo per una regola. *Nota*: il filtro del guadagno cumulativo è disponibile solo per i modelli di associazione creati dopo la release 8.5 o per i modelli precedenti che contengono una misurazione del guadagno cumulativo. I modelli di sequenza non includono questa opzione.

Fare clic su **OK** per applicare i filtri attivati in questa finestra di dialogo.

Generazione di grafici per le regole

I nodi Associazione forniscono un gran numero di informazioni. Ciononostante, non sempre rappresentano un formato facilmente accessibile per gli utenti di business. Per fornire dati che siano facilmente incorporabili nei report di business, nelle presentazioni, e così via, è possibile produrre dei grafici dei dati selezionati. La scheda Modello consente di generare un grafico per una determinata regola, ossia solo per i casi previsti da tale regola.

1. Nella scheda Modello selezionare la regola desiderata.
2. Dal menu Genera scegliere **Grafico (da selezione)**. Viene visualizzata la scheda Di base del nodo Lavagna grafica.

   Nota: quando si visualizza la Lavagna grafica in questo modo, sono disponibili soltanto le schede Di base e Dettagliato.
3. Mediante le impostazioni della scheda Di base o Dettagliato, specificare i dettagli da visualizzare sul grafico.
4. Fare clic su OK per generare il grafico.

L'intestazione del grafico indica la regola e i dettagli antecedenti selezionati per essere inclusi.

Impostazioni del nugget del modello di regole di associazione

La scheda Impostazioni consente di definire le opzioni di calcolo del punteggio per i modelli di associazione (Apriori e CARMA). Questa scheda è disponibile solo dopo che il nugget del modello è stato aggiunto a un flusso per il calcolo del punteggio.

Nota: la finestra di dialogo per la ricerca di un modello grezzo non include la scheda Impostazioni, perché non è possibile calcolarne il punteggio. Per calcolare il punteggio del modello "grezzo", è necessario prima generare un insieme di regole. Per ulteriori informazioni, consultare l'argomento "Generazione di un insieme di regole da un nugget del modello di associazione" a pagina 241.

Numero massimo di previsioni. Specificare il numero massimo di previsioni per ogni insieme di elementi dell'insieme. Questa opzione viene utilizzata insieme a Criterio regola per creare le previsioni "top", in cui *top* indica il livello più elevato di confidenza, supporto, guadagno cumulativo e così via, come specificato di seguito.

Criterio regola. Selezionare la misura che determina l'efficacia delle regole. Le regole sono ordinate in base all'efficacia dei criteri selezionati per restituire le previsioni top per una serie di elementi. Sono disponibili i criteri seguenti:

- Confidenza
- Supporto
- Supporto regola (Supporto * Confidenza)
- Guadagno cumulativo
- Distribuibilità

Consenti ripetizione previsioni. Selezionare questa opzione per includere più regole con lo stesso conseguente durante il calcolo del punteggio. Per esempio, selezionando questa opzione è possibile calcolare il punteggio delle regole seguenti:

```
pane & formaggio -> vino
formaggio & frutta -> vino
```

Deselezionare questa opzione per escludere la ripetizione delle previsioni durante il calcolo del punteggio.

Nota: le regole con più conseguenti (`pane & formaggio & frutta -> vino & patè`) sono considerate previsioni ripetute solo se tutti i conseguenti (`vino & patè`) sono stati precedentemente previsti.

Ignora elementi dell'insieme senza corrispondenza. Selezionare questa opzione per ignorare gli elementi aggiuntivi della serie di elementi. Per esempio, se si seleziona questa opzione per un insieme che contiene `[tenda & sacco a pelo & bollitore]`, la regola `tenda & sacco a pelo -> fornelletto_gas` verrà applicata anche se nell'insieme è presente l'elemento aggiuntivo (`bollitore`).

In alcuni casi, potrebbe essere consigliabile escludere gli elementi aggiuntivi. Per esempio, è probabile che chi acquista una tenda, un sacco a pelo e un bollitore possieda già un fornelletto a gas. e che pertanto quest'ultimo non rappresenti la previsione migliore. In questi casi, è consigliabile deselezionare **Ignora elementi dell'insieme senza corrispondenza** per garantire che gli antecedenti della regola corrispondano esattamente al contenuto dell'insieme. Per default, gli elementi senza corrispondenza vengono ignorati.

Controlla che le previsioni non siano nell'insieme. Selezionare questa opzione per verificare che nell'insieme non siano presenti anche i conseguenti. Per esempio, se si calcola un punteggio allo scopo di definire un'indicazione su prodotti di arredamento, è improbabile che per un insieme che contiene già un tavolo da soggiorno ci sia l'intenzione di acquistarne un altro. In questo caso è opportuno selezionare l'opzione. Nel caso invece di prodotti deperibili o a perdere (per esempio il formaggio, i prodotti per

neonati o i cerotti), le regole che includono già il conseguente nell'insieme potrebbero risultare utili. In quest'ultimo caso, l'opzione migliore è **Non controllare l'insieme per le previsioni**, che viene descritta di seguito.

Controlla che le previsioni siano nell'insieme. Selezionare questa opzione per verificare che nell'insieme siano presenti anche i conseguenti. Questa opzione è utile quando si cerca di ottenere informazioni sui clienti o sulle transazioni esistenti. Per esempio, è possibile identificare le regole con il livello di guadagno cumulativo più alto e quindi individuare i clienti che soddisfano tali regole.

Non controllare l'insieme per le previsioni. Selezionare questa opzione per includere tutte le regole durante il calcolo del punteggio, indipendentemente dalla presenza o meno dei conseguenti nell'insieme.

Riepilogo del nugget del modello di regole di associazione

Nella scheda Riepilogo di un nugget del modello di regole di associazione viene visualizzato il numero di regole scoperte e i valori minimo e massimo per il supporto, il guadagno cumulativo, la confidenza e la distribuibilità delle regole nell'insieme di regole.

Generazione di un insieme di regole da un nugget del modello di associazione

È possibile utilizzare i nugget del modello di associazione quali Apriori e CARMA per calcolare direttamente il punteggio dei dati oppure è possibile generare innanzitutto un sottoinsieme di regole, denominato anche **insieme di regole**. Gli insiemi di regole sono particolarmente utili quando si utilizza il modello grezzo, che non consente di calcolare direttamente il punteggio. Per ulteriori informazioni, consultare l'argomento "Modelli grezzi" a pagina 51.

Per generare un insieme di regole, scegliere **Insieme di regole** dal menu Genera nel browser del nugget del modello. È possibile specificare le opzioni seguenti per la traduzione delle regole in un insieme di regole:

Nome insieme di regole. Consente di specificare il nome del nuovo nodo Insieme di regole generato.

Crea nodo in. Controlla la posizione del nuovo nodo Insieme di regole generato. Selezionare **Area**, **Palette GM** o **Entrambi**.

Campo Obiettivo. Determina quale campo di output verrà utilizzato per il nodo Insieme di regole generato. Selezionare un campo di output singolo dall'elenco.

Supporto minimo. Specifica il supporto minimo per mantenere le regole nell'insieme di regole generato. Le regole con un supporto inferiore al valore specificato non verranno incluse nel nuovo insieme di regole.

Confidenza minima. Specifica la confidenza minima per mantenere le regole nell'insieme di regole generato. Le regole con una confidenza inferiore al valore specificato non verranno incluse nel nuovo insieme di regole.

Valore di default. Consente di specificare un valore di default per il campo obiettivo assegnato ai record di cui è stato calcolato il punteggio e per i quali non viene generata alcuna regola.

Generazione di un modello filtrato

Per generare un modello filtrato da un nugget del modello di associazione (per esempio un nodo Apriori, CARMA o Insieme di regole di sequenza), scegliere **Modello filtrato** dal menu Genera nel browser del nugget del modello. In questo modo verrà creato un modello di sottoinsieme che include solo le regole visualizzate correntemente nel browser. *Nota*: non è possibile generare modelli filtrati per modelli grezzi.

Per il filtro delle regole è possibile specificare le opzioni seguenti:

Nome del nuovo modello. Consente di specificare il nome del nuovo nodo Modello filtrato.

Crea nodo in. Controlla la posizione del nuovo nodo Modello filtrato. Selezionare **Area**, **Palette GM** o **Entrambi**.

Numerazione regole. Specifica la modalità di numerazione degli ID regola nel sottoinsieme di regole inserito nel modello filtrato.

- **Mantieni numeri ID regola originali.** Selezionare questa opzione per mantenere la numerazione originale delle regole. Per default, alle regole viene assegnato un ID corrispondente al relativo ordine di rilevamento da parte dell'algoritmo. Tale ordine può variare in base all'algoritmo utilizzato.
- **Rinumera regole consecutivamente a partire da.** Selezionare questa opzione per assegnare nuovi ID regola alle regole filtrate. I nuovi ID vengono assegnati in base al criterio di ordinamento visualizzato nella tabella del browser delle regole nella scheda Modello, a partire dal numero specificato in questo campo. Per specificare il numero iniziale per gli ID, è possibile utilizzare le frecce a destra.

Calcolo del punteggio delle regole di associazione

I punteggi determinati dall'esecuzione di nuovi dati mediante un nugget del modello di regole di associazione vengono restituiti in campi separati. Per ogni previsione vengono aggiunti tre nuovi campi, dove *P* rappresenta la previsione, *C* rappresenta la confidenza e *I* rappresenta l'ID regola. L'organizzazione di tali campi di output dipende dal formato dei dati di input, transazionale o tabellare. Consultare “Dati in formato tabellare e dati transazionali” a pagina 230 per una panoramica di tali formati.

Si supponga, per esempio, di calcolare il punteggio di dati basket utilizzando un modello che genera previsioni basate sulle tre regole seguenti:

```
Regola_15  pane&vino-> carne (confidenza 54%)
Regola_22  formaggio -> frutta (confidenza 43%)
Regola_5 pane&formaggio -> verdurasurgelata (confidenza 24%)
```

Dati in formato tabellare. Per i dati in formato tabellare, le tre previsioni (3 è il valore di default) vengono restituite in un singolo record.

Tabella 14. Punteggi in formato tabulare.

ID	Pane	Vino	Formaggio	P1	C1	I1	P2	C2	I2	P3	C3	I3
Fred	1	1	1	carne	0.54	15	frutta	0.43	22	verdurasurgelata	24	5

Dati transazionali. Per i dati transazionali, viene generato un record separato per ogni previsione. Le previsioni vengono aggiunte in colonne separate, ma i punteggi vengono restituiti nel momento in cui vengono calcolati. Questo determina record con previsioni incomplete, come mostrato nell'output di esempio riportato di seguito. La seconda e la terza previsione (P2 e P3) sono vuote nel primo record, insieme alle confidenze associate e agli ID regola. Dopo che vengono restituiti i punteggi, il record finale contiene tuttavia le tre previsioni.

Tabella 15. Punteggi in formato transazionale.

ID	Elemento	P1	C1	I1	P2	C2	I2	P3	C3	I3
Fred	pane	carne	0.54	14	$null$	$null$	$null$	$null$	$null$	$null$
Fred	formaggio	carne	0.54	14	frutta	0.43	22	$null$	$null$	$null$
Fred	vino	carne	0.54	14	frutta	0.43	22	verdurasurgelata	24	5

Per includere unicamente previsioni complete per la creazione di report o la distribuzione, è possibile utilizzare un nodo Seleziona per selezionare record completi.

Nota: i nomi dei campi utilizzati in questi esempi sono abbreviati per maggiore chiarezza. Durante l'utilizzo effettivo, i campi dei risultati per i modelli di associazione sono denominati come riportato nella seguente tabella.

Tabella 16. Nomi dei campi dei risultati per i modelli di associazione.

Nuovo campo	**Nome di campo di esempio**
Previsione	*$A-TRANSACTION_NUMBER-1*
Confidenza (o altro criterio)	*$AC-TRANSACTION_NUMBER-1*
ID regola	*$A-Rule_ID-1*

Le regole con più conseguenti

L'algoritmo CARMA consente regole con più conseguenti — ad esempio:

```
pane -> vino&formaggio
```

Quando viene calcolato il punteggio di regole "con più conseguenti", le previsioni vengono restituite nel formato visualizzato nella tabella riportata di seguito.

Tabella 17. Risultati di calcolo del punteggio contenenti una previsione con più conseguenti.

ID	**Pane**	**Vino**	**Formaggio**	**P1**	**C1**	**I1**	**P2**	**C2**	**I2**	**P3**	**C3**	**I3**
Fred	1	1	1	carne&veg	0.54	16	frutta	0.43	22	verdurasurgelata	21	5

In alcuni casi, potrebbe essere necessario suddividere tali punteggi prima della distribuzione. Per suddividere una previsione con più conseguenti, è necessario analizzare il campo mediante le funzioni stringa CLEM.

Deployment dei modelli di associazione

Quando si determina il punteggio di modelli di associazione, previsioni e confidenze vengono inserite in colonne separate (dove *P* rappresenta la previsione, *C* la confidenza e *I* l'ID regola). Questo avviene indipendentemente dal fatto che i dati di input siano in formato tabellare o transazionali. Per ulteriori informazioni, consultare l'argomento "Calcolo del punteggio delle regole di associazione" a pagina 242.

Quando si preparano punteggi per la distribuzione, è possibile che sia necessario trasporre i dati di output in un formato con previsioni disposte in righe anziché in colonne (una previsione per riga, talvolta noto come formato "di registro incrementale acquisti").

Trasposizione di punteggi tabulari

È possibile trasporre punteggi tabulari da colonne a righe utilizzando una combinazione di passaggi in IBM SPSS Modeler, come indicato nella procedura seguente.

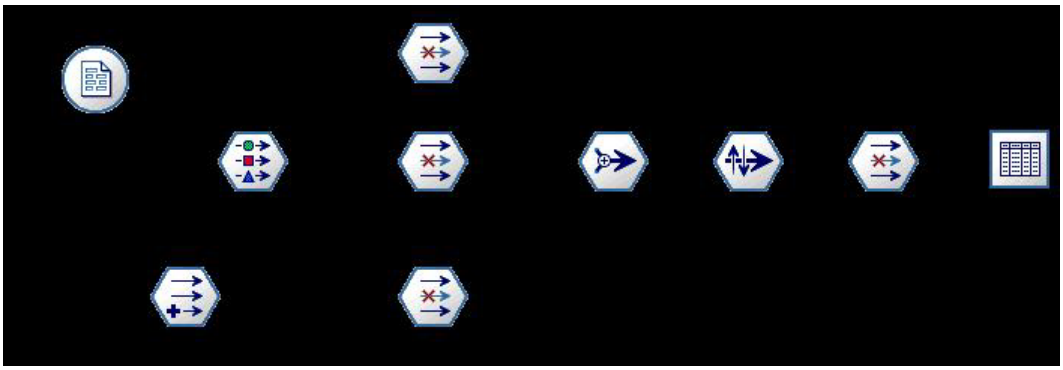


Figura 49. Esempio di flusso utilizzato per trasporre i dati in formato tabellare nel formato "registro incrementale acquisti"

1. Utilizzare la funzione `@INDEX` in un nodo Deriva per verificare l'ordine corrente delle previsioni e salvare l'indicatore in un nuovo campo, per esempio *Ordine_originale*.
2. Aggiungere un nodo tipologia per verificare se tutti i campi sono istanziati.
3. Utilizzare un nodo Filtro per rinominare i campi di default relativi alla previsione, alla confidenza e all'ID (*P1*, *C1*, *I1*) in campi comuni, per esempio *Pred*, *Crit* e *ID_regola*, che verranno utilizzati per accodare record successivamente. Sarà necessario un nodo Filtro per ogni previsione generata.

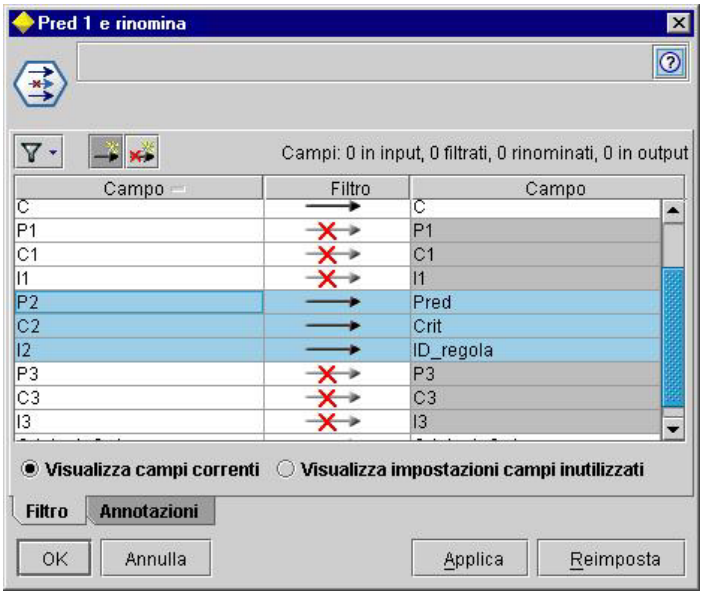


Figura 50. Filtro dei campi per le previsioni 1 e 3 e ridenominazione dei campi per la previsione 2.

4. Utilizzare un nodo Accodamento per accodare i valori per i campi condivisi *Pred*, *Crit* e *ID_regola*.
5. Collegare un nodo Ordina per ordinare in ordine crescente i record del campo *Ordine_originale* e in ordine decrescente i record del campo*Crit*, che viene utilizzato per ordinare le previsioni in base al criterio, per esempio la confidenza, il guadagno cumulativo e il supporto.
6. Utilizzare un altro campo Filtro per filtrare il campo *Ordine_originale* dall'output.

A questo punto i dati sono pronti per la distribuzione.

Trasposizione di punteggi transazionali

Il processo è simile per la trasposizione dei punteggi transazionali. Per esempio, il flusso riportato di seguito traspone punteggi in un formato con un'unica previsione per riga, così come richiesto per la distribuzione.

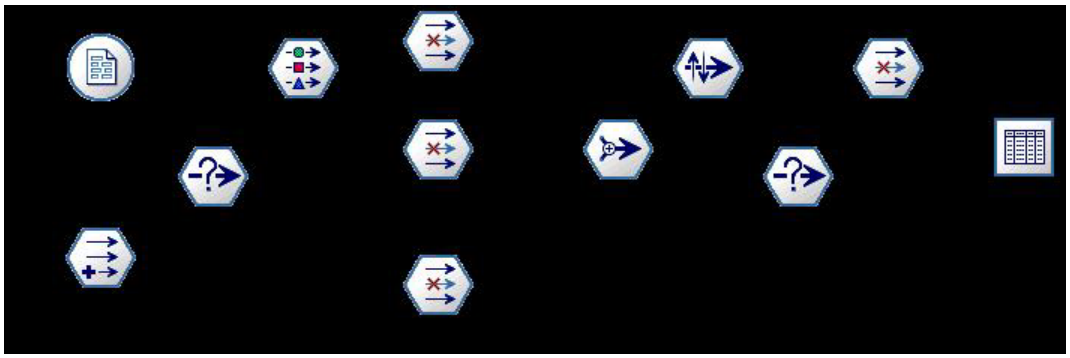


Figura 51. Esempio di flusso utilizzato per la trasposizione di dati transazionali nel formato "registro incrementale acquisti"

Eccetto che per l'aggiunta di due nodi Seleziona, il processo è identico a quello descritto in precedenza per i dati in formato tabellare.

- Il primo nodo Seleziona consente di confrontare gli ID regola di record adiacenti e di includere solo i record univoci o non definiti. Questo nodo Seleziona utilizza l'espressione CLEM per selezionare i record: `ID /= @OFFSET(ID,-1) or @OFFSET(ID,-1) = undef`.
- Il secondo nodo Seleziona consente di scartare le regole estranee oppure le regole nelle quali `ID_regola` ha un valore null. Questo nodo Seleziona utilizza la seguente espressione CLEM per scartare i record: `not(@NULL(Rule_ID))`.

Per ulteriori informazioni sulla trasposizione dei punteggi per la distribuzione, rivolgersi al supporto tecnico.

Nodo Sequenza

Il nodo Sequenza consente di individuare gli schemi nei dati sequenziali o basati su valori temporali, nel formato `pane -> formaggio`. Gli elementi di una sequenza sono **serie di elementi** che costituiscono una singola transazione. Per esempio, se una persona si reca in negozio e compra pane e latte e dopo alcuni giorni torna per comprare del formaggio, la sua attività di acquisto può essere rappresentata come due serie di elementi. La prima serie di elementi contiene il pane e il latte e il secondo contiene il formaggio. Per **sequenza** si intende un elenco di serie di elementi che tendono a ricorrere secondo un ordine prevedibile. Mediante il nodo Sequenza è possibile rilevare sequenze frequenti e creare un nodo di modello generato utilizzabile per elaborare previsioni.

Requisiti. Per creare un insieme di regole Sequenza, è necessario specificare un campo ID, un campo ora facoltativo e uno o più campi contenuto. Si noti che queste impostazioni devono essere definite nella scheda Campi del nodo Sequenza; non possono essere lette da un nodo Tipo posto a monte. Il campo ID può avere qualsiasi ruolo o livello di misurazione. Se si specifica un campo ora, il campo può avere qualsiasi ruolo ma l'archiviazione deve essere di tipo numerico, data, ora o timestamp. Se non si specifica un campo ora, il nodo Sequenza utilizzerà un timestamp implicito, servendosi di fatto di numeri di riga come valori temporali. I campi contenuto possono avere qualsiasi ruolo e livello di misurazione, ma tutti i campi contenuto devono essere dello stesso tipo. Se sono numerici, devono essere intervalli di numeri interi (non intervalli di numeri reali).

Efficacia. Il nodo Sequenza si basa sull'algoritmo delle regole di associazione CARMA, che utilizza un metodo efficiente in due passaggi per trovare le sequenze. Inoltre, il nodo del modello generato creato da

un nodo Sequenza può essere inserito in un flusso di dati per creare delle previsioni. Il nodo del modello generato può anche generare Supernodi per rilevare e contare sequenze specifiche e per eseguire previsioni basate su sequenze specifiche.

Opzioni dei campi del nodo Sequenza

Prima di eseguire un nodo Sequenza, è necessario specificare i campi ID e contenuto nella scheda Campi del nodo Sequenza. Se si desidera utilizzare un campo ora, è necessario specificarlo qui.

Campo ID. Selezionare un campo ID dall'elenco. Come campo ID è possibile utilizzare campi numerici o simbolici. Ogni valore univoco di questo campo deve indicare una specifica unità di analisi. In un'applicazione per analisi market basket, per esempio, ciascun ID potrebbe rappresentare un singolo cliente. Per un'applicazione per analisi di registri Web, ciascun ID potrebbe rappresentare un computer (per indirizzo IP) o un utente (per dati di login).

- **ID contigui.** Se i dati sono preordinati in modo che tutti i record con lo stesso ID vengano raggruppati 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 Sequenza ordinerà i dati automaticamente.

Nota: se i propri dati non sono ordinati e viene selezionata questa opzione, è possibile che vengano visualizzati risultati non validi nel modello Sequenza.

Campo ora. Se nei dati si desidera utilizzare un campo per indicare le ore degli eventi, selezionare **Utilizza campo ora** e specificare il campo da utilizzare. Il campo ora deve essere di tipo numerico, data, ora o timestamp. Se non si specifica nessun campo ora, si presume che i record arrivino dalla sorgente dati in ordine sequenziale e come valori di ora vengono utilizzati i numeri di record (il primo record arriva all'ora `"1"`, il secondo all'ora `"2"` e così via).

Campi contenuto. Specificare il campo o i campi contenuto per il modello. Questi campi contengono gli eventi rilevanti nella creazione di modelli di sequenza.

Il nodo Sequenza può gestire dati sia in formato tabulare che transazionale. Se si utilizzano più campi multipli con dati transazionali, si presume che gli elementi specificati in questi campi per un particolare record rappresentino gli elementi trovati in una singola transazione con un singolo timestamp. Per ulteriori informazioni, consultare l'argomento “Dati in formato tabellare e dati transazionali” a pagina 230.

Partizione. Questo campo consente di specificare un campo utilizzato per partizionare i dati in campioni distinti per le fasi di addestramento, di test e di convalida della creazione del modello. Utilizzando un campione per generare il modello e un altro campione per sottoporlo a verifica, è possibile ottenere un'indicazione valida del modo in cui il modello potrà essere esteso a insiemi di dati di dimensioni maggiori, simili ai dati correnti. Se sono stati definiti più campi partizione utilizzando nodi Tipo o Partizione, 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. Si noti inoltre che per applicare la partizione selezionata nell'analisi, è necessario che il partizionamento sia attivato anche nella scheda Opzioni modello relativa al nodo. La disattivazione di questa opzione consente di disattivare il partizionamento senza dover modificare le impostazioni dei campi.

Opzioni del modello di nodo Sequenza

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.

Utilizza dati partizionati. Se è definito un campo partizione, questa opzione garantisce che per la creazione del modello vengano utilizzati dati solo dalla partizione di addestramento.

Supporto minimo regola (%). È possibile specificare un criterio di supporto. Per **supporto regola** si intende la proporzione di ID nei dati di addestramento che contengono l'intera sequenza. Per prendere in esame le frequenze più comuni, aumentare il valore di questa impostazione.

Confidenza minima regola (%). È possibile specificare un criterio di confidenza per mantenere le sequenze nell'insieme sequenza. Per **confidenza** si intende la percentuale di ID in cui viene eseguita una previsione corretta, tra tutti gli ID per i quali la regola fa una previsione. La confidenza viene calcolata come numero di ID per i quali viene trovata l'intera sequenza, diviso per il numero di ID per i quali si trovano gli antecedenti, in base ai dati di addestramento. Le sequenze con una confidenza minore rispetto al criterio specificato vengono scartate. Se si ottengono troppe sequenze o sequenze non interessanti, provare ad aumentare il valore di questa impostazione. Se le sequenze ottenute sono troppo poche, provare a diminuire il valore di questa impostazione.

Dimensione massima sequenza. È possibile impostare il numero massimo di *serie di elementi* distinte (in contrapposizione agli *elementi*) in una sequenza. Se le sequenze interessanti sono relativamente brevi, è possibile diminuire il valore di questa impostazione per accelerare la creazione dell'insieme di sequenze.

Previsioni da aggiungere al flusso. Specificare il numero di previsioni che il nodo Modello generato risultante deve aggiungere al flusso. Per ulteriori informazioni, consultare l'argomento "Nugget del modello Sequenza" a pagina 248.

Opzioni avanzate del nodo Sequenza

Le seguenti opzioni avanzate consentono agli esperti del nodo Sequenza di ottimizzare il processo di generazione dei modelli. Per accedere alle opzioni avanzate, impostare la modalità su **Livello avanzato** nella scheda Livello avanzato.

Imposta durata massima. Se questa opzione è selezionata, le sequenze verranno limitate a quelle con una durata (il tempo tra la prima e l'ultima serie di elementi) inferiore o uguale al valore specificato. Se non è stato specificato un campo ora, la durata viene espressa in termini di righe (record) nei dati non elaborati. Se il campo temporale utilizzato è un campo ora, data o timestamp, la durata è espressa in secondi. Per i campi numerici, la durata è espressa nelle stesse unità del campo.

Imposta valore di taglio. Durante l'elaborazione, l'algoritmo CARMA utilizzato nel nodo Sequenza rimuove periodicamente (**taglia**) le serie di elementi non frequenti dall'elenco delle serie di elementi potenziali per preservare la memoria. Selezionare questa opzione per modificare la frequenza delle operazioni di taglio. Il numero specificato determina la frequenza di taglio. Specificare un valore minore per diminuire i requisiti di memoria dell'algoritmo (aumentando tuttavia il tempo di addestramento richiesto), oppure specificare un valore maggiore per accelerare l'addestramento (tenendo tuttavia presente che questa impostazione può aumentare i requisiti di memoria).

Imposta numero massimo di sequenze in memoria. Se questa opzione è selezionata, durante la creazione del modello l'algoritmo CARMA limiterà la memorizzazione delle sequenze candidate al numero di sequenze specificato. Selezionare questa opzione se IBM SPSS Modeler utilizza troppa memoria durante la generazione dei modelli Sequenza. Si noti che il valore massimo delle sequenze specificato qui è il numero di sequenze candidate di cui si tiene traccia internamente durante la generazione del modello. Questo numero deve essere molto più grande del numero di sequenze che si prevede di ottenere nel modello finale.

Limita intervalli tra serie di elementi. Questa opzione permette di specificare i vincoli in termini di intervalli di tempo che separano le serie di elementi. Se è selezionata, le serie di elementi con intervalli di tempo inferiori al valore **Intervallo minimo** o maggiori del valore **Intervallo massimo** specificati non verranno considerate parte di una sequenza. Utilizzare questa opzione per evitare di contare sequenze che includono lunghi intervalli di tempo o quelle che si verificano in un intervallo molto breve.

Nota: se il campo ora utilizzato è un campo ora, data o timestamp, l'intervallo di tempo è espresso in secondi. Per i campi numerici, l'intervallo di tempo è espresso nelle stesse unità dell'ora.

Ad esempio, considerare il seguente elenco di transazioni.

Tabella 18. Elenco di transazioni di esempio.

ID	Ora	Sommario
1001	1	mele
1001	2	pane
1001	5	formaggio
1001	6	condimento

Se si genera un modello in base a questi dati con l'intervallo minimo impostato su 2, si ottengono le seguenti sequenze:

```
mele -> formaggio

mele -> condimento

pane -> formaggio

pane -> condimento
```

Non si otterrebbero sequenze quali `mele -> pane`, perché l'intervallo tra `mele` e `pane` è inferiore all'intervallo minimo. Allo stesso modo, considerare i dati alternativi riportati di seguito.

Tabella 19. Elenco di transazioni di esempio.

ID	Ora	Sommario
1001	1	mele
1001	2	pane
1001	5	formaggio
1001	20	condimento

Se l'intervallo massimo fosse impostato su 10, non verrebbe visualizzata alcuna sequenza con `condimento`, perché l'intervallo tra `formaggio` e `condimento` è troppo grande perché siano considerati parte della stessa sequenza.

Nugget del modello Sequenza

I nugget del modello Sequenza rappresentano le sequenze rilevate per un determinato campo di output tramite il nodo Sequenza ed è possibile aggiungerli a flussi per generare previsioni.

Quando si esegue un flusso contenente un nodo Sequenza, tale nodo aggiunge una coppia di campi contenenti previsioni e valori di confidenza associati per ogni previsione dal modello Sequenza ai dati. Per default, vengono aggiunte tre coppie di campi contenenti le tre previsioni principali (e i valori di confidenza associati). È possibile modificare il numero di previsioni generate quando viene generato il modello impostando le opzioni del modello del nodo Sequenza al momento della generazione oppure nella scheda Impostazioni dopo aver aggiunto il nugget del modello a un flusso. Per ulteriori informazioni, consultare l'argomento "Impostazioni del nugget del modello Sequenza" a pagina 252.

I nomi dei nuovi campi derivano dal nome del modello. I nomi dei campi sono *$S-sequenza-n* per il campo della previsione (dove *n* indica l' *n*esima previsione) e *$SC-sequenza-n* per il campo della confidenza. In un flusso con più nodi Regole di sequenza in una serie, i nomi dei nuovi campi

includeranno dei numeri nel prefisso come criterio di differenziazione. Il primo nodo Insieme di sequenza nel flusso utilizzerà la denominazione standard, il secondo nodo aggiungerà i prefissi *$S1-* e *$SC1-*, il terzo i prefissi *$S2-* e *$SC2-* e così via. Le previsioni vengono visualizzate in ordine di confidenza in modo tale che *$S-sequenza-1* contenga la previsione con la più alta confidenza, *$S-sequenza-2* contenga la previsione con la seconda più alta confidenza e così via. Nei record caratterizzati da un numero di previsioni disponibili inferiore al numero di previsioni richieste, le previsioni restanti contengono il valore `$null$`. Per esempio, se è possibile eseguire solo due previsioni per un determinato record, i valori di *$S-sequenza-3* e *$SC-sequenza-3* saranno `$null$`.

Per ogni record, le regole del modello vengono confrontate con l'insieme di transazioni elaborate per l'ID corrente fino al momento attuale, incluso il record corrente e tutti i record precedenti con lo stesso ID e un timestamp precedente. Le regole *k* con i valori di confidenza più elevati che vengono applicate a questo insieme di transazioni vengono utilizzate per generare le previsioni *k* per il record, dove *k* indica il numero di previsioni specificato nella scheda Impostazioni dopo aver aggiunto il modello al flusso. (Se più regole prevedono lo stesso risultato per l'insieme di transazioni, viene utilizzata solo la regola con il valore di confidenza più elevato). Per ulteriori informazioni, consultare l'argomento “Impostazioni del nugget del modello Sequenza” a pagina 252.

Analogamente ad altri tipi di modelli di regole di associazione, il formato dei dati deve corrispondere al formato utilizzato durante la generazione del modello Sequenza. Per esempio, è possibile utilizzare i modelli generati utilizzando dati in formato tabellare per determinare il punteggio esclusivamente di dati in formato tabellare. Per ulteriori informazioni, consultare l'argomento “Calcolo del punteggio delle regole di associazione” a pagina 242.

Nota: quando viene calcolato il punteggio dei dati utilizzando un nodo Insieme di sequenza in un flusso, le impostazioni relative alla tolleranza o all'intervallo selezionate durante la creazione del modello vengono ignorate per motivi di calcolo del punteggio.

Previsioni da regole di sequenza

Il nodo gestisce i record in modo diverso a seconda dell'ora (o in base all'ordine, se non è stato utilizzato alcun campo timestamp per creare il modello). I record devono essere ordinati in base ai campi ID e timestamp (se presente). Tuttavia, le previsioni non sono legate al timestamp del record al quale vengono aggiunte. Si riferiscono semplicemente agli elementi che hanno più probabilità di verificarsi *in un determinato momento del futuro*, data la cronologia delle transazioni relative all'ID corrente fino al record corrente.

Si tenga presente che le previsioni per ogni record non dipendono necessariamente dalle transazioni di tale record. Se le transazioni del record corrente non generano una regola specifica, le regole verranno selezionate in base alle transazioni precedenti per l'ID corrente. In altre parole, se il record corrente non aggiunge informazioni utili per la previsione alla sequenza, al record corrente viene applicata la previsione dell'ultima transazione utile per l'ID.

Si supponga per esempio di avere un modello Sequenza con l'unica regola

```
Marmellata -> Pane (0.66)
```

e di passare i seguenti record.

Tabella 20. Record di esempio.

ID	Acquisto	Previsione
001	marmellata	pane
001	latte	pane

Si noti che il primo record genera una previsione di *pane* , come ci si aspetta. Anche il secondo record contiene una previsione di *pane*, perché non esiste alcuna regola per *marmellata* seguita da *latte*; quindi, la transazione *latte* non aggiunge alcuna informazioni utile e resta valida la regola `Marmellata -> Pane`.

Generazione di nuovi nodi

Il menu Genera consente di creare nuovi Supernodi basati sul modello Sequenza.

- **Supernodo regola.** Crea un Supernodo in grado di rilevare e contare le occorrenze delle sequenze nei punteggi dei dati. Questa opzione non è attiva se non viene selezionata una regola. Per ulteriori informazioni, consultare l'argomento "Generazione di un Supernodo regola da un nugget del modello Sequenza" a pagina 252.
- **Modello a palette.** Restituisce il modello nella palette dei modelli. Può rivelarsi utile in quelle situazioni in cui un collega abbia inviato un flusso che contiene il modello ma non il modello stesso.

Dettagli del nugget del modello Sequenza

Nella scheda Modello di un nugget del modello Sequenza sono visualizzate le regole estratte dall'algoritmo. Ogni riga della tabella rappresenta una regola, con l'antecedente (la parte "se" della regola) nella prima colonna seguito dal conseguente (la parte "allora" della regola) nella seconda colonna.

Ciascuna regola è mostrata nel seguente formato.

Tabella 21. Formato della regola

Antecedente	Conseguente
`birra e verdura_scatola`	`birra`
`pesce` `pesce`	`pesce`

La prima regola di esempio è interpretata come *per gli ID con "birra" e "verdura in scatola" nella stessa transazione, è probabile che l'occorrenza successiva sia "birra."* La seconda regola di esempio può essere interpretata come *per gli ID con "pesce" in una transazione e "pesce" in un'altra, è probabile che l'occorrenza successiva sia "pesce."* Si noti che nella prima regola *birra* e *verdura_scatola* vengono acquistate contemporaneamente, mentre, nella seconda regola, *pesce* è acquistato in due transazioni distinte.

Menu Ordina. Il pulsante del menu Ordina sulla barra degli strumenti controlla l'ordinamento delle regole. La direzione dell'ordinamento (crescente o decrescente) può essere modificata utilizzando il pulsante di direzione dell'ordinamento (freccia in su o in giù).

È possibile ordinare le regole in base a:

- % supporto
- % confidenza
- % supporto regola
- Conseguente
- Primo antecedente
- Ultimo antecedente
- Numero di elementi (antecedenti)

Per esempio, la tabella seguente è ordinata in ordine decrescente in base al numero di elementi: Regole con più elementi nell'insieme di elementi antecedente precedono quelle con un numero inferiore di elementi.

Tabella 22. Regole ordinate per numero di elementi

Antecedente	Conseguente
birra e verdura_scatola e carne_surgelata	carne_surgelata
birra e verdura_scatola	birra
pesce pesce	pesce
softdrink	softdrink

Menu Mostra/Nascondi criteri. Il pulsante del menu Mostra/Nascondi criteri (icona griglia) controlla le opzioni di visualizzazione delle regole. Sono disponibili le seguenti opzioni di visualizzazione:

- **Istanze** visualizza le informazioni relative al numero di ID univoci per cui si verifica la *sequenza completa* — antecedenti e conseguenti —. Si noti la differenza con i modelli Associazione, per i quali il numero di istanze si riferisce al numero di ID che includono *solo* gli antecedenti. Per esempio, data la regola `pane -> formaggio`, il numero di ID dei dati di addestramento che includono sia *pane* che *formaggio* è definito come **istanze**.
- **Supporto** visualizza la proporzione di ID nei dati di addestramento per la quale gli antecedenti sono veri. Per esempio, se il 50% dei dati di addestramento include l'antecedente *pane*, il supporto per la regola `pane -> formaggio` sarà pari al 50%. A differenza dei modelli Associazione, il supporto *non* si basa sul numero di istanze, come indicato in precedenza.
- **Confidenza** visualizza la percentuale di ID in cui viene eseguita una previsione corretta, tra tutti gli ID per i quali la regola fa una previsione. La confidenza viene calcolata come numero di ID per i quali viene trovata l'intera sequenza, diviso per il numero di ID per i quali si trovano gli antecedenti, in base ai dati di addestramento. Per esempio, se il 50% dei dati di addestramento contiene `verdura_scatola` (a indicare il supporto antecedente) ma solo il 20% contiene sia `verdura_scatola` che `carne_surgelata`, allora la confidenza per la regola `verdura_scatola -> carne_surgelata` sarà `Supporto regola/Supporto antecedente` oppure, in questo caso, il 40%.
- **Supporto regola** per i modelli Sequenza è basato su istanze e visualizza la proporzione dei record di addestramento per i quali l'intera regola, ovvero gli antecedenti e i conseguenti, sono veri. Ad esempio, se il 20% dei dati di addestramento contiene sia il *pane* sia il *formaggio*, il supporto regola per la regola `pane -> formaggio` è pari al 20%.

Si noti che le proporzioni si basano sulle transazioni valide (transazioni con almeno un elemento osservato o valore vero) anziché sulle transazioni totali. Le transazioni non valide — senza elementi o valori veri — vengono ignorate per questi calcoli.

Pulsante Filtro. Il pulsante Filtro (icona a forma di imbuto) sul menu espande la parte inferiore della finestra di dialogo e visualizza un riquadro che contiene i filtri della regola attivi. I filtri consentono di limitare il numero delle regole visualizzate nella scheda Modelli.

Figura 52. Pulsante Filtro

Per creare un filtro, fare clic sull'icona Filtro a destra del riquadro espanso. Verrà aperta una finestra di dialogo in cui è possibile specificare i vincoli per la visualizzazione delle regole. Si noti che il pulsante Filtro viene spesso utilizzato insieme al menu Genera per filtrare innanzitutto le regole e creare quindi un modello che contiene tale sottoinsieme di regole. Per ulteriori informazioni, consultare “Definizione dei filtri per le regole” a pagina 238 di seguito.

Impostazioni del nugget del modello Sequenza

Nella scheda Impostazioni di un nugget del modello Sequenza sono visualizzate le opzioni di calcolo del punteggio del modello. Questa scheda è disponibile solo dopo che il modello è stato aggiunto alle aree del flusso per il calcolo del punteggio.

Numero massimo di previsioni. Specificare il numero massimo di previsioni per ogni insieme di elementi dell'insieme. Le regole con i valori di confidenza più elevati che vengono applicate a questo insieme di transazioni sono utilizzate per generare previsioni per il record fino al limite specificato.

Scheda Riepilogo del nugget del modello Sequenza

Nella scheda Riepilogo di un nugget del modello di regole di sequenza viene visualizzato il numero di regole scoperte e i valori minimo e massimo per il supporto e la confidenza delle regole. Se è stato eseguito un nodo Analisi collegato a questo nodo Modelli, in questa sezione verranno visualizzate anche le informazioni sui risultati di tale analisi.

Per ulteriori informazioni, consultare l'argomento “Esplorazione dei nugget del modello” a pagina 42.

Generazione di un Supernodo regola da un nugget del modello Sequenza

Per generare un Supernodo regola in base a una regola di sequenza:

1. Nella scheda Modello di un nugget del modello di regole di sequenza, fare clic su una riga della tabella per selezionare la regola desiderata.
2. Dai menu del browser di regole, scegliere:

   Genera > **Supernodo regola**

Importante: per utilizzare il Supernodo generato, è necessario ordinare i dati per campo ID (e campo Ora, se presente), prima di passarli al Supernodo. Il Supernodo non rileverà le sequenze correttamente nei dati non ordinati.

È possibile specificare le opzioni seguenti nella generazione di un Supernodo regola:

Rileva. Specifica la definizione delle corrispondenze per i dati passati nel Supernodo.

- **Solo antecedenti.** Il Supernodo identificherà una corrispondenza ogni volta che trova gli antecedenti relativi alla regola selezionata nell'ordine corretto all'interno di un insieme di record aventi lo stesso ID, indipendentemente dal fatto che anche il conseguente sia stato rilevato o meno. Si noti che tale opzione non prende in considerazione le impostazioni relative al vincolo dell'intervallo degli elementi o della tolleranza timestamp del nodo Sequenza originale. Quando la serie di elementi relativa all'ultimo antecedente viene rilevata nel flusso (e tutti gli altri antecedenti sono stati rilevati nell'ordine corretto), tutti i record successivi aventi l'ID corrente includeranno il riepilogo selezionato di seguito.
- **Intera sequenza.** Il SuperNodo identificherà una corrispondenza ogni volta che trova gli antecedenti e il conseguente per la regola selezionata nell'ordine corretto all'interno di un insieme di record aventi lo stesso ID. Tale opzione non prende in considerazione le impostazioni relative al vincolo dell'intervallo degli elementi o della tolleranza timestamp del nodo Sequenza originale. Quando viene rilevato il conseguente nel flusso (e anche tutti gli antecedenti sono stati rilevati nell'ordine corretto), tutti i record successivi aventi l'ID corrente includeranno il riepilogo selezionato di seguito.

Visualizzazione. Controlla come vengono aggiunti i riepiloghi corrispondenti ai dati nell'output del Supernodo Regola.

- **Valore conseguente per la prima occorrenza.** Il valore aggiunto ai dati è il valore del conseguente previsto basato sulla prima occorrenza della corrispondenza. I valori vengono aggiunti come un campo nuovo denominato *rule_n_consequent*, dove *n* è il numero della regola (basato sull'ordine di creazione dei Supernodi Regola nel flusso).

- **Valore vero per la prima occorrenza.** Il valore aggiunto ai dati è vero se esiste almeno una corrispondenza relativa all'ID, falso se non esiste alcuna corrispondenza. I valori vengono aggiunti come un campo nuovo denominato *regola_n_flag*.
- **Conteggio delle occorrenze.** Il valore aggiunto ai dati è il numero di corrispondenze relative all'ID. I valori vengono aggiunti come un campo nuovo denominato *regola_n_conteggio*.
- **Numero di regola.** Il valore aggiunto è il numero della regola selezionata. I **numeri di regola** vengono assegnati in base all'ordine in cui viene aggiunto il Supernodo al flusso. Per esempio, il Supernodo Regola della prima regola viene considerato come *regola 1*, il Supernodo Regola della seconda regola come *regola 2* e così via. Questa opzione è molto utile quando si includono più Supernodi Regola nel flusso. I valori vengono aggiunti come un campo nuovo denominato *regola_n_numero*.
- **Includi cifre di confidenza.** Se selezionata, questa opzione aggiunge la confidenza della regola al flusso dei dati nonché il riepilogo selezionato. I valori vengono aggiunti come un campo nuovo denominato *regola_n_confidenza*.

Capitolo 13. Modelli di serie temporali

Perché si effettuano le previsioni

Effettuare una previsione significa prevedere i valori di una o più serie nel tempo. Per esempio, potrebbe essere utile prevedere la domanda per una linea di prodotti o servizi al fine di assegnare le risorse necessarie per la produzione o la distribuzione. Poiché la pianificazione delle decisioni richiede tempo, le previsioni rappresentano uno strumento essenziale in molti processi di pianificazione.

I metodi delle serie temporali di modellazione suppongono che la storia si ripeta — se non esattamente, in modo sufficientemente simile da consentire di prendere decisioni migliori grazie allo studio del passato. Per esempio, per prevedere le vendite per l'anno successivo si potrebbe iniziare ad analizzare le vendite dell'anno corrente, procedendo all'indietro per individuare eventuali tendenze o schemi ricorrenti che si sono sviluppati negli ultimi anni. Gli schemi, tuttavia, possono essere difficili da determinare. Se le vendite aumentano per molte settimane di fila, per esempio, ci si potrebbe chiedere se il fenomeno fa parte di un ciclo stagionale o se sia invece l'inizio di una tendenza a lungo termine.

Con le tecniche di modellazione statistica è possibile analizzare gli schemi presenti nei dati passati ed effettuarne una proiezione per determinare un intervallo all'interno del quale è probabile che rientrino i valori futuri della serie. Il risultato è una previsione più precisa su cui basare le proprie decisioni.

Dati di serie temporali

Una **serie temporale** è una raccolta ordinata di misurazioni effettuate a intervalli regolari, quali per esempio i corsi azionari giornalieri o i dati delle vendite settimanali. Le misurazioni possono riguardare qualunque elemento e ogni serie può in genere essere classificata nelle seguenti categorie:

- **Dipendente.** Una serie per cui si desidera effettuare una previsione.
- **Predittore.** Una serie che può contribuire a spiegare l'obiettivo, per esempio l'utilizzo di un budget pubblicitario per prevedere le vendite. I predittori possono essere utilizzati solo con i modelli ARIMA.
- **Evento.** Una serie predittore speciale utilizzata per rappresentare incidenti ricorrenti prevedibili — ad esempio, le promozioni sulle vendite.
- **Intervento.** Una serie predittore speciale utilizzata per spiegare eventi isolati passati, per esempio un black-out o uno sciopero dei dipendenti.

Gli intervalli possono rappresentare qualsiasi unità di tempo, ma l'intervallo deve essere uguale per tutte le misurazioni. Inoltre, tutti gli intervalli per cui non esistono misurazioni devono essere impostati sul valore mancante. Pertanto, il numero degli intervalli per cui esistono delle misurazioni (compresi quelli con valori mancanti) definisce la lunghezza dell'intervallo storico dei dati.

Caratteristiche delle serie temporali

Lo studio del comportamento passato di una serie è utile per individuare schemi ricorrenti ed effettuare precisioni migliori. Quando vengono rappresentate in un grafico, molte serie temporali presentano una o più delle seguenti caratteristiche:

- Tendenze
- Cicli stagionali e non stagionali
- Impulsi e passaggi
- Anomali

Tendenze

Una **tendenza** è uno spostamento graduale del livello della serie verso l'alto o verso il basso, oppure la tendenza dei valori della serie ad aumentare o diminuire nel tempo.

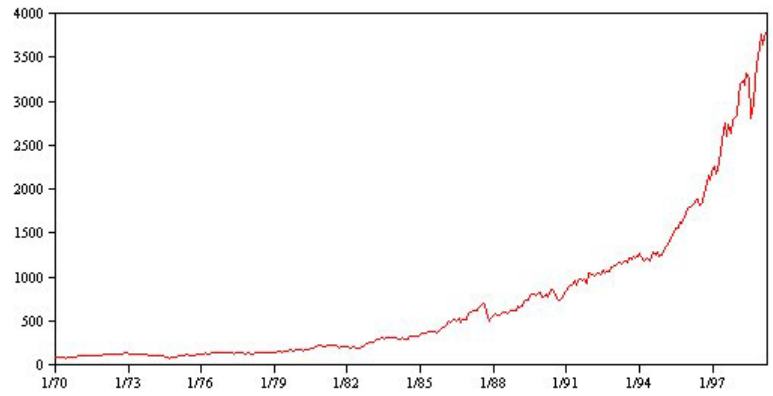


Figura 53. Tendenza

Le tendenze possono essere **locali** o **globali**, ma in una sola serie possono essere presenti entrambi i tipi di tendenza. Storicamente, le rappresentazioni grafiche delle serie relative all'indice di un mercato azionario mostrano una tendenza globale al rialzo. In periodi di recessione sono state rilevate tendenze locali al ribasso e tendenze locali al rialzo in periodi di prosperità.

Le tendenze possono inoltre essere **lineari** o **non lineari**. Le tendenze lineari sono incrementi additivi positivi o negativi del livello della serie, paragonabili all'effetto dell'interesse semplice sul capitale. Le tendenze non lineari sono spesso moltiplicative, con incrementi proporzionali ai valori delle serie precedenti.

Le tendenze lineari globali sono adeguate e consentono di effettuare previsioni corrette sia mediante modelli ARIMA che di livellamento esponenziale. Nella creazione dei modelli ARIMA, le serie che evidenziano delle tendenze vengono in genere differenziate per eliminare l'effetto della tendenza.

Cicli stagionali

Un **ciclo stagionale** è uno schema ripetitivo e prevedibile nei valori della serie.

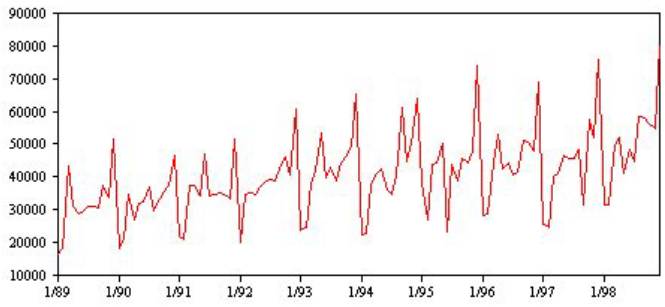


Figura 54. Ciclo stagionale

I cicli stagionali sono legati all'intervallo della serie presa in esame. Per esempio, generalmente i dati mensili sono caratterizzati da cicli trimestrali e annuali. Una serie mensile potrebbe evidenziare un ciclo trimestrale significativo con un calo nel primo trimestre oppure un ciclo annuale potrebbe mostrare un picco ogni mese di dicembre. Si ritiene che le serie che mostrano un ciclo stagionale evidenzino una **stagionalità**.

Gli schemi stagionali sono utili per ottenere adattamenti e previsioni corretti ed esistono modelli ARIMA e di livellamento esponenziale in grado di rilevare la stagionalità.

Cicli non stagionali

Un **ciclo non stagionale** è uno schema ripetitivo e talvolta imprevedibile nei valori della serie.

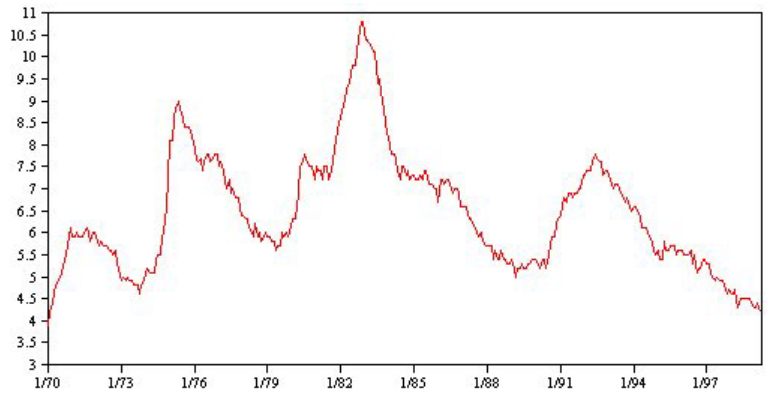


Figura 55. Ciclo non stagionale

Alcune serie, quali il tasso di disoccupazione, mostrano chiaramente un comportamento ciclico; tuttavia, la periodicità del ciclo può variare nel tempo e rendere difficile prevedere quando si verificherà un aumento o una diminuzione. Altre serie possono avere cicli prevedibili ma non essere adattabili al calendario gregoriano, oppure avere cicli più lunghi di un anno. Per esempio, le maree seguono il calendario lunare, il traffico merci e passeggeri legato alle Olimpiadi aumenta ogni quattro anni ed esistono numerose ricorrenze religiose le cui date nel calendario gregoriano cambiano di anno in anno.

Gli schemi ciclici non stagionali sono difficili da modellare e in genere aumentano l'incertezza delle previsioni. Per esempio, nel mercato azionario vi sono numerosi esempi di serie che è stato impossibile prevedere, malgrado gli sforzi degli analisti. Ciononostante, qualora esistano, gli schemi non stagionali devono essere spiegati. In molti casi è possibile comunque individuare un modello sufficientemente adatto ai dati cronologici, che offre la migliore possibilità di ridurre al minimo l'incertezza della previsione.

Impulsi e passaggi

In molte serie si verificano bruschi cambiamenti di livello, che generalmente sono di due tipi:

- Un cambiamento nel livello della serie improvviso e *temporaneo*, o **impulso**
- Un cambiamento nel livello della serie improvviso e *permanente*, o **passaggio**

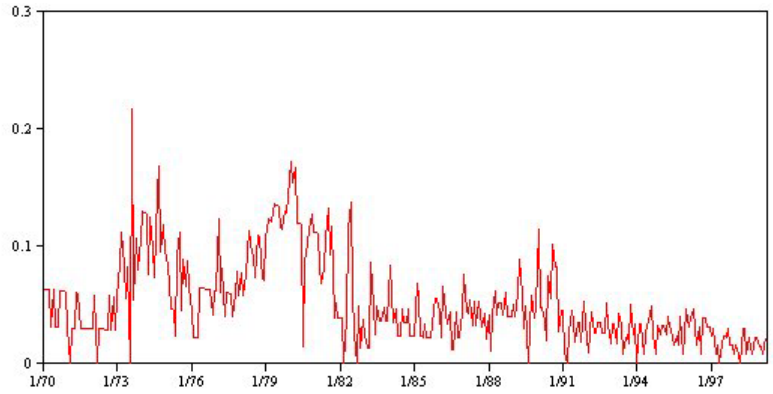


Figura 56. Serie con un impulso

Quando vengono osservati passaggi o impulsi, è importante trovare una spiegazione plausibile. I modelli di serie temporali sono studiati per spiegare i cambiamenti graduali, non quelli improvvisi. Di conseguenza, essi tendono a sottovalutare gli impulsi e a essere compromessi dai passaggi, il che determina uno scarso adattamento dei modelli e previsioni incerte (è possibile che alcuni casi di stagionalità evidenzino cambiamenti di livello improvvisi, ma il livello rimane costante da un periodo stagionale all'altro).

Se può essere spiegata, un'interferenza può essere modellata mediante un **intervento** o un **evento**. Per esempio, nell'agosto del 1973 un embargo sul petrolio imposto dall'OPEC, l'organizzazione dei paesi esportatori di petrolio, provocò una brusca variazione del tasso d'inflazione, che tornò ai livelli normali nei mesi successivi. Specificando l'**intervento di un punto** per il mese dell'embargo è possibile migliorare l'adattamento del modello e quindi, indirettamente, le previsioni. Per esempio, un punto vendita al dettaglio potrebbe scoprire che le vendite sono state decisamente superiori al normale nei giorni in cui tutti gli articoli erano scontati del 50%. Specificando la promozione con lo sconto del 50% come **evento** ricorrente è possibile migliorare l'adattamento del modello e stimare l'effetto di una ripetizione della promozione in futuro.

Anomali

I cambiamenti di livello di una serie temporale che non è possibile spiegare vengono definiti **valori anomali**. Queste osservazioni sono incoerenti rispetto al resto della serie e possono incidere considerevolmente sull'analisi e, di conseguenza, influire sulla capacità di previsione del modello di serie temporali.

La figura che segue mostra vari tipi di valori anomali rilevati di frequente nelle serie temporali. Le righe blu rappresentano una serie priva di valori anomali, mentre le righe rosse suggeriscono lo schema che potrebbe essere presente se la serie contenesse dei valori anomali. Questi valori anomali sono classificati tutti come **deterministici** perché influiscono solo sul livello medio della serie.

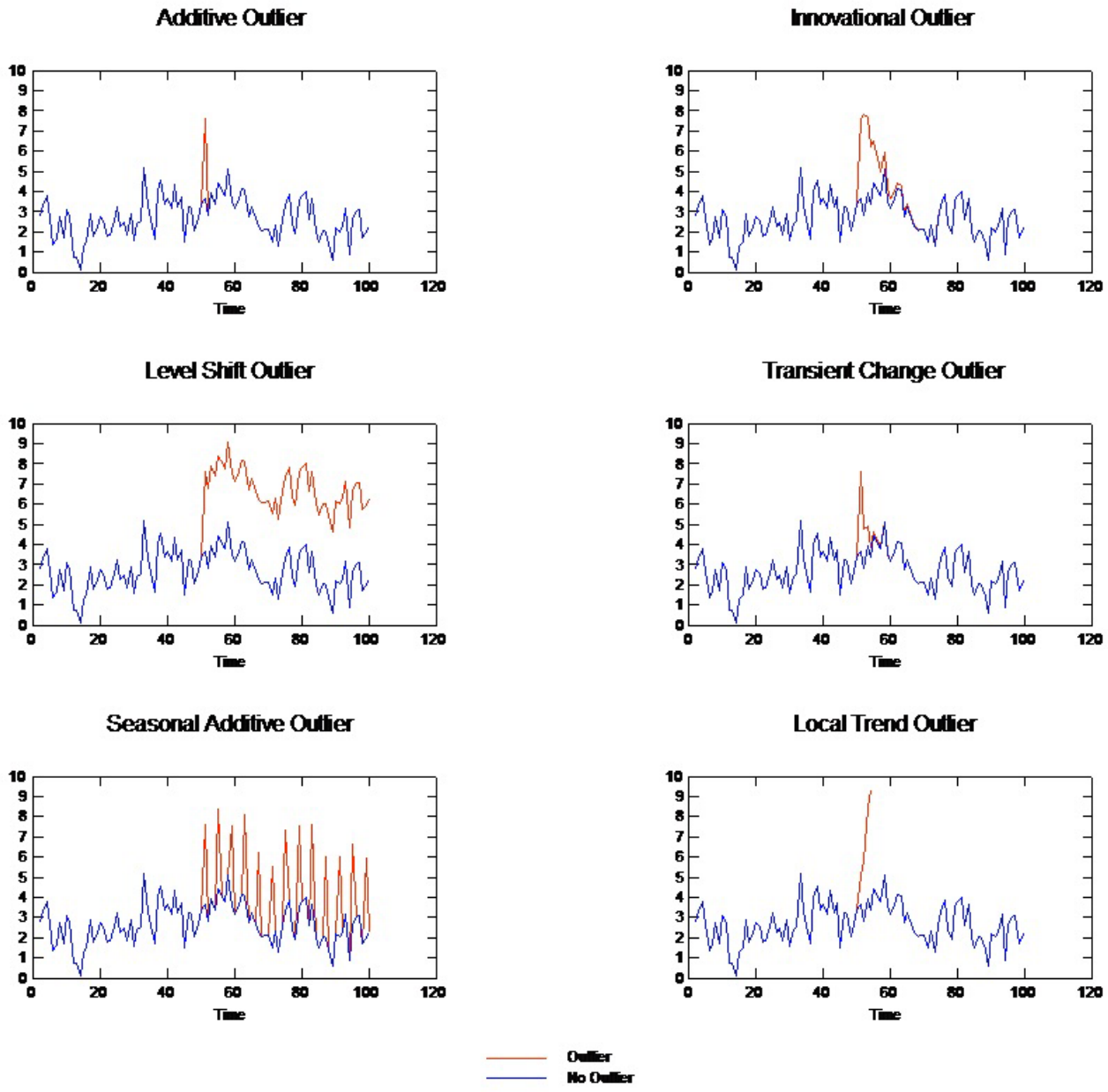


Figura 57. Tipi di valori anomali

- **Valore anomalo additivo.** Un valore anomalo additivo si presenta come un valore straordinariamente grande o piccolo riscontrato per una singola osservazione. Le osservazioni seguenti non risentono dell'effetto di un valore anomalo additivo. I valori anomali additivi consecutivi vengono generalmente definiti **patch di valori anomali additivi**.
- **Valore anomalo innovativo.** Un valore anomalo innovativo è caratterizzato da un impatto iniziale con effetti che si riscontrano anche nelle osservazioni successive. L'influenza dei valori anomali può aumentare con il passare del tempo.
- **Valore anomalo di cambiamento di livello.** Per un cambiamento di livello, tutte le osservazioni che compaiono dopo il valore anomalo si spostano su un nuovo livello. A differenza dei valori anomali additivi, i valori anomali di cambiamento di livello influiscono su molte osservazioni e hanno un effetto permanente.

- **Valore anomalo di variazione transiente.** I valori anomali di variazione transiente sono simili ai valori anomali di cambiamento di livello, ma l'effetto del valore anomalo diminuisce esponenzialmente nelle osservazioni successive. Alla fine, la serie torna al livello normale.
- **Valore anomalo additivo stagionale.** Un valore anomalo additivo stagionale si presenta come un valore straordinariamente grande o piccolo che ricorre ripetutamente a intervalli regolari.
- **Valore anomalo di tendenza locale.** Un valore anomalo di tendenza locale provoca una fluttuazione generale nella serie causata da uno schema nei valori anomali dopo l'insorgenza del valore anomalo iniziale.

Il rilevamento dei valori anomali nelle serie temporali comporta l'individuazione della posizione, del tipo e dell'ampiezza di tutti i valori anomali presenti. Tsay (1988) ha proposto una procedura iterativa per rilevare la variazione del livello medio al fine di individuare i valori anomali deterministici. Questo processo comporta il confronto di un modello di serie temporali che presuppone l'assenza di valori anomali con un altro modello che include dei valori anomali. Le differenze fra i due modelli forniscono stime dell'effetto del trattamento di un determinato punto come valore anomalo.

Funzioni di autocorrelazione e autocorrelazione parziale

L'autocorrelazione e l'autocorrelazione parziale sono misure dell'associazione tra i valori della serie corrente e quelli delle serie passate che indicano quali sono i valori delle serie passate più utili per prevedere valori futuri. Conoscendo questo dato è possibile determinare l'ordine dei processi in un modello ARIMA. Più precisamente:

- **Funzione di autocorrelazione (ACF).** Al ritardo *k*, questa è la correlazione tra i valori della serie che sono separati da *k* intervalli.
- **Funzione di autocorrelazione parziale (PACF).** Al ritardo *k*, questa è la correlazione tra i valori della serie che sono separati da *k* intervalli, tenuto conto dei valori degli intervalli intermedi.

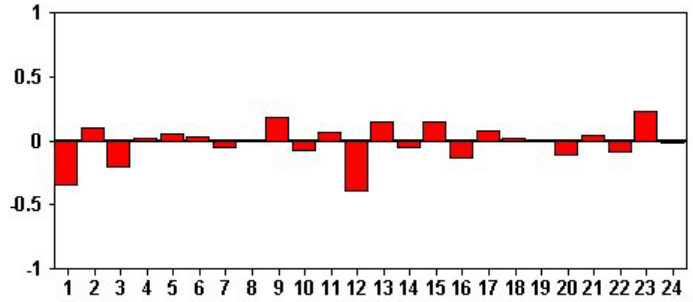


Figura 58. Plot ACF di una serie

L'asse *x* del grafico ACF indica il ritardo con cui viene calcolata l'autocorrelazione; l'asse *y* indica il valore della correlazione (tra −1 e 1). Per esempio, un picco in corrispondenza del ritardo 1 in un plot ACF indica una forte correlazione tra ogni valore della serie e il valore precedente, un picco in corrispondenza del ritardo 2 indica una forte correlazione tra ogni valore e il valore che ricorre due punti prima, e così via.

- Una correlazione positiva indica che i valori correnti maggiori corrispondono ai valori maggiori in corrispondenza del ritardo specificato; una correlazione negativa indica che i valori correnti maggiori corrispondono ai valori minori in corrispondenza del ritardo specificato.
- Il valore assoluto di una correlazione è la misura dell'intensità dell'associazione: i valori assoluti più grandi indicano le relazioni più forti.

Trasformazioni di serie

Le trasformazioni risultano spesso utili per stabilizzare una serie prima di effettuare la stima dei modelli. Questo è importante soprattutto per i modelli ARIMA, che richiedono che la serie sia **stazionaria** prima

di procedere alla stima dei modelli. Una serie è stazionaria se il livello globale (media) e la deviazione media dal livello (varianza) sono costanti in tutta la serie.

Benché gran parte delle serie interessanti non siano stazionarie, i modelli ARIMA sono efficaci a condizione che sia possibile rendere la serie stazionaria applicando trasformazioni quali il logaritmo naturale, la differenziazione o la differenziazione stagionale.

Trasformazioni per la stabilizzazione della varianza. Le serie in cui la varianza cambia nel tempo si possono in molti casi stabilizzare mediante la trasformazione logaritmica naturale o la trasformazione radice quadrata, note anche come trasformazioni funzionali.

- **Log naturale.** Il logaritmo naturale viene applicato ai valori della serie.
- **Radice quadrata.** La funzione radice quadrata viene applicata ai valori della serie.

Le trasformazioni logaritmica naturale e radice quadrata non possono essere utilizzate per le serie con valori negativi.

Trasformazioni per la stabilizzazione del livello. Una lenta diminuzione dei valori dell'ACF indica che ogni valore della serie ha una forte correlazione con il valore precedente. Analizzando la variazione dei valori della serie si ottiene un livello stabile.

- **Differenziazione semplice.** Vengono calcolate le differenze fra ogni valore della serie e quello precedente, escludendo il valore più vecchio della serie. Le serie differenziate avranno quindi un valore in meno rispetto alle serie originali.
- **Differenziazione stagionale.** Identica alla differenziazione semplice, tranne per il fatto che vengono calcolate le differenze tra i singoli valori e i valori stagionali precedenti.

Quando si utilizzano la differenziazione semplice o quella stagionale contemporaneamente alla trasformazione logaritmica o radice quadrata, la trasformazione per la stabilizzazione della varianza viene sempre applicata per prima. Quando si utilizza sia la differenziazione semplice che quella stagionale, i valori della serie risultante sono uguali indipendentemente dal fatto che venga applicata prima la differenziazione semplice o quella stagionale.

Serie predittore

Le serie predittore comprendono dati correlati che possono contribuire a spiegare il comportamento della serie da prevedere. Per esempio, un rivenditore al dettaglio tramite Web o catalogo potrebbe prevedere le vendite in base al numero di cataloghi spediti, al numero di linee telefoniche disponibili o al numero di risultati di ricerca ottenuti dalla pagina Web della società.

Qualunque serie può essere utilizzata come predittore, a condizione che si estenda nel futuro quanto la previsione da effettuare e disponga di dati completi, senza valori mancanti.

Prestare attenzione quando si aggiungono dei predittori a un modello. L'aggiunta di un gran numero di predittori aumenterà il tempo necessario per la stima dei modelli. Sebbene l'aggiunta di predittori possa migliorare la capacità di adattamento del modello ai dati cronologici, questo non significa necessariamente che il modello sia in grado di effettuare una previsione migliore, per cui potrebbe non valere la pena di aumentare la complessità del processo. In teoria, l'obiettivo dovrebbe essere l'individuazione del modello più semplice in grado di effettuare previsioni corrette.

Come regola generale, si consiglia di utilizzare un numero di predittori inferiore alle dimensioni del campione diviso 15 (al massimo un predittore ogni 15 casi).

Predittori con dati mancanti. I predittori con dati mancanti o incompleti non possono essere utilizzati per le previsioni. Questo vale sia per i dati cronologici che per i valori futuri. In alcuni casi è possibile evitare questa limitazione impostando l'intervallo di stima del modello in modo da escludere i dati più vecchi dalla stima dei modelli.

Nodo Modelli Serie temporali

Il nodo Serie temporali stima i modelli di livellamento esponenziale, i modelli ARIMA (Autoregressive Integrated Moving Average, autoregressivi integrati a media mobile) univariati e ARIMA (o a funzione di trasferimento) multivariati per le serie temporali e genera previsioni basate sui dati di serie temporali.

Il **livellamento esponenziale** è un metodo di previsione che utilizza i valori ponderati delle osservazioni di serie precedenti per prevedere i valori futuri. In senso stretto, il livellamento esponenziale non si basa su un'analisi teorica dei dati, ma effettua la previsione di un punto alla volta, modificando le previsioni di pari passo con l'acquisizione di nuovi dati. Questa tecnica è utile per la previsione di serie che evidenziano tendenza, stagionalità o entrambe le caratteristiche. È possibile scegliere tra vari modelli di livellamento esponenziale che si differenziano tra loro per il trattamento della tendenza e della stagionalità.

I modelli ARIMA costituiscono un metodo più sofisticato per la modellazione dei componenti di tendenza e di stagionalità rispetto ai modelli di livellamento esponenziale e, in particolare, offrono in più il vantaggio di includere nel modello variabili (predittore) indipendenti. Ciò comporta l'indicazione esplicita di ordini autoregressivi e di media mobile, nonché del grado di differenziazione. È possibile includere variabili predittore e definire funzioni di trasferimento per tutte o alcune di esse e specificare il rilevamento automatico dei valori anomali o di un insieme specifico di valori anomali.

Nota: in termini pratici, i modelli ARIMA sono utili soprattutto se si desidera includere dei predittori che possono contribuire a spiegare il comportamento della serie oggetto della previsione, quale il numero di cataloghi inviati per posta o il numero di risultati di ricerca ottenuti per la pagina Web di una società. I modelli di livellamento esponenziale descrivono il comportamento della serie temporale senza cercare di capire le ragioni di tale comportamento. Per esempio, è probabile che una serie che storicamente ha raggiunto il punto massimo ogni 12 mesi continui con il medesimo andamento, anche se le ragioni di tale comportamento sono sconosciute.

È inoltre disponibile la funzione **Expert Modeler**, che tenta di identificare e stimare automaticamente il modello ARIMA o di livellamento esponenziale più adatto per una o più variabili obiettivo, eliminando così la necessità di individuare un modello appropriato procedendo per tentativi. In caso di dubbi, utilizzare Expert Modeler.

Se vengono specificate delle variabili predittore, per l'inclusione nei modelli ARIMA Expert Modeler seleziona le variabili aventi una relazione statisticamente significativa con la serie dipendente. Laddove è opportuno, le variabili del modello vengono trasformate tramite la differenziazione e/o una trasformazione a radice quadrata o logaritmica naturale. Per default, Expert Modeler considera tutti i modelli ARIMA e di livellamento esponenziale e sceglie quello migliore per ogni campo obiettivo. È possibile tuttavia impostare Expert Modeler in modo che scelga solo il modello di livellamento esponenziale migliore o il modello ARIMA migliore. È inoltre possibile specificare il rilevamento automatico dei valori anomali.

Esempio. Un analista di una società che fornisce accesso a banda larga su scala nazionale deve generare una previsione relativa agli abbonamenti stipulati dagli utenti al fine di prevedere l'utilizzo della larghezza di banda. È necessario effettuare una previsione per ognuno dei mercati locali che costituiscono la base nazionale degli abbonati. È possibile utilizzare i modelli di serie temporali per eseguire le previsioni per i tre mesi successivi per una serie di mercati locali.

Requisiti

Il nodo Serie temporali è diverso dagli altri nodi di IBM SPSS Modeler perché non è possibile inserirlo semplicemente in un flusso ed eseguire il flusso. Il nodo Serie temporali deve sempre essere preceduto da un nodo Intervalli di tempo che specifica informazioni quali l'intervallo di tempo da utilizzare (anni, trimestri, mesi, ecc.), i dati da usare per la stima ed eventualmente l'arco di tempo coperto dalla previsione.

I dati di serie temporali devono essere ripartiti 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 di cui si dispone non possiedono già queste caratteristiche, il nodo Intervalli di tempo è in grado di trasformarli nel formato richiesto.

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

Opzioni dei campi

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 Serie temporali utilizza le informazioni sui campi da un nodo Tipo situato 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. 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 di serie temporali

Nome modello. Specifica il nome assegnato al modello generato quando viene eseguito il nodo.

- **Auto.** Genera il nome del modello automaticamente in base ai nomi dei campi ID e Obiettivo oppure il nome del tipo di modello nei casi in cui l'obiettivo non viene specificato (come i modelli cluster).
- **Personalizzato.** Consente di specificare un nome personalizzato per il nugget del modello.

Continua valutazione mediante modelli esistenti. Se è già stato generato un modello di serie temporali, selezionare questa opzione per riutilizzare i criteri impostati per quel modello e generare un nuovo nodo del modello nella palette Modelli piuttosto che ricominciare la procedura di creazione dall'inizio. In questo modo è possibile risparmiare tempo effettuando una nuova stima e generando una nuova previsione basata sulle stesse impostazioni del modello precedente, ma utilizzando dati più recenti. Così, per esempio, se il modello originale di una determinata serie temporale era la tendenza lineare di Holt, viene utilizzato lo stesso tipo di modello per effettuare una nuova stima e una previsione; il sistema non tenta di individuare nuovamente il tipo di modello più adatto per i nuovi dati. Selezionando questa opzione si disattivano i controlli di **Metodo** e **Criteri**. Per ulteriori informazioni, consultare l'argomento “Esecuzione di una nuova stima e previsione” a pagina 269.

Metodo. Le opzioni disponibili sono Expert Modeler, Livellamento esponenziale o ARIMA. Per ulteriori informazioni, consultare l'argomento “Nodo Modelli Serie temporali” a pagina 262. Selezionare **Criteri** per specificare le opzioni per il metodo selezionato.

- **Expert Modeler.** Selezionare questa opzione per utilizzare Expert Modeler, che individua automaticamente il modello più adatto per ogni serie dipendente.
- **Livellamento esponenziale.** Utilizzare questa opzione per specificare un modello di livellamento esponenziale personalizzato.
- **ARIMA.** Utilizzare questa opzione per specificare un modello ARIMA personalizzato.

Informazioni sugli intervalli di tempo

Questa sezione della finestra di dialogo contiene informazioni sulle specifiche per la stima e la previsione effettuate sul nodo Intervalli di tempo. Si noti che questa sezione non viene visualizzata se si sceglie l'opzione **Continua la stima mediante i modelli esistenti**.

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 determinerà un errore quando si tenta di eseguire il flusso; a monte del nodo Serie temporali è necessario infatti includere un nodo Intervalli di tempo.

Informazioni varie

Larghezza limite di confidenza (%). Gli intervalli di confidenza vengono calcolati per i valori stimati del modello e le autocorrelazioni dei residui. È possibile specificare qualsiasi valore positivo inferiore a 100. Per impostazione predefinita, si utilizza un intervallo di confidenza al 95%.

Numero massimo di ritardi nell'output di ACF e PACF. È possibile impostare il numero massimo di ritardi mostrato nelle tabelle e nei grafici delle autocorrelazioni e delle autocorrelazioni parziali.

Solo modello di calcolo del punteggio di creazione. Selezionare questa casella per ridurre la quantità di dati archiviata nel modello. Questo può migliorare le prestazioni quando vengono creati modelli con grandi quantità (decine di migliaia) di serie temporali. Se si seleziona questa opzione, le schede Modello, Parametri e Residui non vengono visualizzate nel nugget del modello Serie temporali, ma è comunque possibile calcolare il punteggio dei dati nel modo consueto.

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 considererà solo la regressione semplice anziché 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). Per ulteriori informazioni, consultare l'argomento "Impulsi e passaggi" a pagina 257.

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 "Anomali" a pagina 258.

Criteri di livellamento esponenziale per le serie temporali

Tipo di modello. I modelli di livellamento esponenziale vengono classificati come stagionali o non stagionali [1]. I modelli stagionali sono disponibili solo se la periodicità definita mediante il nodo Intervalli di tempo 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, $p = 12$.

1. Gardner, E. S. 1985. Exponential smoothing: The state of the art. *Journal of Forecasting,* 4, 1-28.

- **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" a pagina 260.

- **Nessuna.** Non viene eseguita nessuna trasformazione.
- **Radice quadrata.** Viene eseguita la trasformazione radice quadrata.
- **Log naturale.** Viene eseguita la trasformazione logaritmica naturale.

Criteri ARIMA per le serie temporali

Il nodo Serie temporali 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

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.

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" a pagina 260.

- **Nessuna.** Non viene eseguita nessuna trasformazione.
- **Radice quadrata.** Viene eseguita la trasformazione radice quadrata.
- **Log naturale.** Viene eseguita la trasformazione logaritmica naturale.

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 è visualizzata solo se sono specificati campi di input (con il ruolo impostato su *Input*), nel nodo Tipo oppure nella scheda Campi del nodo Serie temporali (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 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.

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)
- Innovazionale
- Transiente
- Additivo stagionale
- Tendenza locale
- Patch additiva

Per ulteriori informazioni, consultare l'argomento “Anomali” a pagina 258.

Generazione di modelli di serie temporali

Questa sezione fornisce informazioni generali su alcuni aspetti della generazione di modelli di serie temporali:

- Generazione di più modelli
- Utilizzo dei modelli di serie temporali nelle previsioni
- Esecuzione di una nuova stima e previsione

Il nugget del modello generato è descritto in un argomento a parte. Per ulteriori informazioni, consultare l'argomento “Nugget del modello di serie temporali” a pagina 270.

3. Pena, D., G. C. Tiao, and R. S. Tsay, eds. 2001. *A course in time series analysis*. New York: John Wiley and Sons.

Generazione di più modelli

La creazione di modelli di serie temporali in IBM SPSS Modeler genera un singolo modello (ARIMA o di livellamento esponenziale) per ogni campo obiettivo. Quindi, se i campi obiettivo sono più di uno, IBM SPSS Modeler genera più modelli in una sola operazione, risparmiando tempo e consentendo il confronto delle impostazioni dei singoli modelli.

Se si desidera confrontare un modello ARIMA e un modello di livellamento esponenziale per il medesimo campo obiettivo, è possibile eseguire separatamente il nodo Serie temporali specificando ogni volta un modello diverso.

Utilizzo dei modelli di serie temporali nelle previsioni

L'operazione di creazione di una serie temporale utilizza una serie specifica di casi ordinati, nota come intervallo di stima, per creare un modello utilizzabile per prevedere i valori futuri della serie. Questo modello contiene informazioni sul periodo di tempo utilizzato, compreso l'intervallo. Per effettuare previsioni con questo modello è necessario utilizzare gli stessi dati relativi al periodo di tempo e all'intervallo con la stessa serie sia per la variabile obiettivo, sia per le variabili predittore.

Si supponga per esempio di voler prevedere all'inizio di gennaio le vendite mensili del Prodotto 1 per i primi tre mesi dell'anno. Va creato un modello utilizzando i dati effettivi delle vendite mensili per il Prodotto 1 da gennaio a dicembre dell'anno precedente (che chiameremo Anno 1), impostando l'intervallo di tempo su "Mesi". Il modello così creato si può utilizzare per prevedere le vendite del Prodotto 1 per i primi tre mesi dell'Anno 2.

In realtà sarebbe possibile effettuare previsioni per il numero di mesi futuri desiderato, ma più la previsione è spostata nel futuro, meno efficace diventa il modello. Non sarebbe invece possibile effettuare previsioni per le prime tre settimane dell'Anno 2, perché l'intervallo utilizzato per creare il modello è "Mesi". Inoltre, non avrebbe senso utilizzare questo modello per prevedere le vendite del Prodotto 2 - un modello di serie temporale è rilevante solo per i dati utilizzati per definirlo.

Esecuzione di una nuova stima e previsione

Nel modello generato, il periodo di stima è impostato come hardcode. Pertanto, se si applica il modello corrente a nuovi dati, tutti i valori al di fuori del periodo di stima vengono ignorati. Ne consegue che per un modello di serie temporali è necessario effettuare una nuova stima ogni volta che sono disponibili dati nuovi, diversamente da quanto accade per gli altri modelli di IBM SPSS Modeler che possono invece essere riapplicati senza modifiche ai fini del calcolo del punteggio.

Per riprendere l'esempio precedente, si supponga di disporre, all'inizio di aprile nell'Anno 2, dei dati relativi alle vendite mensili effettive da gennaio a marzo. Se si riapplica il modello generato all'inizio di gennaio, la previsione sarà effettuata ancora da gennaio a marzo e i dati effettivi relativi alle vendite di quel periodo saranno ignorati.

La soluzione consiste nel generare un nuovo modello basato sui dati effettivi aggiornati. Presupponendo di non modificare i parametri di previsione, il nuovo modello si può utilizzare per prevedere le vendite dei tre mesi successivi, da aprile a giugno. Se è ancora possibile accedere al flusso utilizzato per generare il modello originale, è sufficiente sostituire il riferimento al file di origine in quel flusso con un riferimento al file contenente i dati aggiornati e rieseguire il flusso per generare il nuovo modello. Se invece si dispone solo del modello originale salvato come file, è possibile comunque utilizzarlo per generare un nodo Serie temporali che può essere quindi aggiunto a un nuovo flusso contenente un riferimento al file di origine aggiornato. A condizione che questo nuovo flusso preceda il nodo Serie temporali con un nodo Intervalli di tempo in cui l'intervallo è impostato su "Mesi", l'esecuzione di questo nuovo flusso genererà il nuovo modello richiesto.

Nugget del modello di serie temporali

L'operazione di modellazione delle serie temporali crea una serie di nuovi campi con il prefisso $TS- come illustrato nella seguente tabella.

Tabella 23. Nuovi campi creati dall'operazione di modellazione delle serie temporali.

Nome del campo	Descrizione
$TS-*nome colonna*	Il valore previsto dal modello per ogni serie obiettivo.
$TSLCI-*nome colonna*	Gli intervalli di confidenza inferiori per ogni serie prevista.*
$TSUCI-*nome colonna*	Gli intervalli di confidenza superiori per ogni serie prevista.*
$TSNR-*nome colonna*	Il valore dei residui di rumore per ogni colonna dei dati del modello generato.*
$TS-Totale	Il totale dei valori della colonna $TS-*nomecolonna* per questa riga.
$TSLCI-Totale	Il totale dei valori della colonna $TSLCI-*nomecolonna* per questa riga.*
$TSUCI-Totale	Il totale dei valori della colonna $TSUCI-*nomecolonna* per questa riga.*
$TSNR-Totale	Il totale dei valori della colonna $TSNR-*nomecolonna* per questa riga.*

* La visibilità di questi campi (per esempio, nell'output di un nodo Tabella collegato) dipende dalle opzioni selezionate sulla scheda Impostazioni del nugget del modello di serie temporali. Per ulteriori informazioni, consultare l'argomento "Impostazioni del modello di serie temporali" a pagina 273.

Il nugget del modello di serie temporali visualizza i dettagli dei vari modelli selezionati per ogni serie inserita nel nodo di creazione Serie temporali. È possibile inserire più serie (per esempio dati relativi a linee di prodotti, aree geografiche o punti vendita) e per ogni serie obiettivo viene generato un modello separato. Per esempio, se le entrate nell'area geografica Est si adattano a un modello ARIMA mentre quelle dell'area geografica Ovest si adattano solo a una media mobile semplice, il punteggio di ogni area geografica viene calcolato con il modello appropriato.

Per ogni modello creato, l'output di default mostra il tipo di modello, il numero di predittori specificati e la misura della bontà di adattamento (*R*-quadrato stazionaria è il valore di default). Se sono stati specificati metodi per i valori anomali, è presente una colonna che mostra il numero di valori anomali rilevati. L'output di default comprende anche colonne per i dati relativi alla statistica Q di Ljung-Box, ai gradi di libertà e ai valori di significatività.

È possibile scegliere anche l'output avanzato, che mostra le seguenti colonne aggiuntive:

- R-quadrato
- RMSE (radice errore quadratico medio)
- MAPE (errore assoluto medio percentuale)
- MAE (errore assoluto della media)
- MaxAPE (errore assoluto percentuale massimo)
- MaxAE (errore assoluto massimo)
- BIC normalizzato (criterio informativo bayesiano normalizzato)

Genera. Consente di generare un nodo Modelli Serie temporali nel flusso o un nugget del modello nella palette.

- **Genera nodo modellazione.** Inserisce in un flusso un nodo Modelli Serie temporali con le impostazioni utilizzate per creare l'insieme di modelli. Questa operazione può essere utile, per esempio, se si dispone di un flusso in cui si desidera utilizzare le impostazioni di questi modelli ma non si dispone più del nodo Modelli utilizzato per generarli.
- **Modello a palette.** Inserisce un nugget del modello contenente tutti gli obiettivi nel manager Modelli.

Modello

Figura 59. Pulsanti Seleziona tutto e Deseleziona tutto

Caselle di controllo. Scegliere i modelli da utilizzare nel calcolo del punteggio. Per default sono selezionate tutte le caselle. I pulsanti **Seleziona tutto** e **Deseleziona tutto** agiscono su tutte le caselle contemporaneamente.

Ordina per. Consente di ordinare le righe di output in ordine crescente o decrescente in una colonna specifica della visualizzazione. L'opzione "Selezionati" ordina l'output in base a una o più righe selezionate con le caselle di controllo. Questa operazione può essere utile, per esempio, per visualizzare i campi obiettivo da "Mercato_1" a "Mercato_9" prima di "Mercato_10", poiché il criterio di ordinamento di default visualizza "Mercato_10" subito dopo "Mercato_1".

Visualizza. La visualizzazione di default (Semplice) mostra l'insieme di base delle colonne di output. L'opzione Avanzate visualizza delle colonne aggiuntive per le misure della bontà di adattamento.

Numero di record utilizzati nella stima. Il numero di righe del file di dati di origine originale.

Obiettivo. Il campo o i campi identificati come obiettivo (con ruolo *Obiettivo*) nel nodo Tipo.

Modello. Il tipo di modello utilizzato per il campo obiettivo.

Predittori. Il numero di predittori (con ruolo *Input*) utilizzato per il campo obiettivo.

Valori anomali. Questa colonna è visualizzata solo se è stato richiesto il rilevamento automatico dei valori anomali in Expert Modeler o tra i criteri ARIMA. Il valore visualizzato è il numero dei valori anomali rilevati.

R-quadrato stazionario. Una misura che confronta la parte stazionaria del modello con un modello di media semplice. Questa misura è preferibile all'R-quadrato semplice se è presente una tendenza o un motivo stagionale. L'R-quadrato stazionario è negativo con un intervallo negativo infinito tendente a 1. I valori negativi indicano che il modello esaminato è peggiore del modello di base. I valori positivi indicano che il modello esaminato è migliore del modello di base.

R-quadrato. Misura della bontà dell'adattamento di un modello lineare, detto anche coefficiente di determinazione. È la proporzione di variabilità della variabile dipendente spiegata dal modello di regressione. Può variare tra 0 e 1. Valori bassi indicano che il modello non si adatta bene ai dati.

RMSE. Acronimo di Root Mean Square Error. La radice dell'errore quadratico medio. Indica di quanto una serie dipendente varia rispetto al livello stimato del modello. Questa misura viene espressa nelle stesse unità della serie dipendente.

MAPE. Errore percentuale assoluto medio. Indica di quanto la serie dipendente varia rispetto al livello predittivo del modello. L'errore è indipendente dalle unità usate e può quindi essere usato per confrontare le serie con unità diverse.

MAE. Errore assoluto della media. Misura di quanto la serie varia rispetto al livello predittivo del modello. Il valore MAE viene indicato in unità originali della serie.

MaxAPE. Errore percentuale assoluto massimo. Errore stimato più alto, espresso in percentuale. Questa misura è utile per stimare lo scenario peggiore delle previsioni.

MaxAE. Errore assoluto massimo. L'errore stimato più alto, espresso nelle stesse unità della serie dipendente. Al pari di MaxAPE, è utile per stimare lo scenario peggiore per le previsioni. L'errore assoluto massimo e l'errore percentuale assoluto massimo possono verificarsi in punti diversi della serie; ad esempio se l'errore assoluto di una serie grande è leggermente maggiore dell'errore assoluto del valore minore della serie. In questo caso l'errore assoluto massimo si verifica nel valore maggiore della serie mentre l'errore percentuale assoluto massimo si verifica nel valore minore della serie.

BIC normalizzato. Criterio di informazione bayesiano normalizzato. Misura generale dell'adattamento generale di un modello che tenta di tenere conto della complessità del modello. Questa misura è un punteggio basato sull'errore quadratico medio e comprende una penalità per il numero di parametri del modello e una lunghezza per la serie. La penalità rimuove il vantaggio dei modelli con più parametri, consentendo di confrontare più facilmente la statistica con più modelli della stessa serie.

Q. La statistica Q di Ljung-Box. Si tratta di un test della casualità degli errori residui in questo modello.

df. Gradi di libertà. Il numero di parametri del modello che possono variare durante la stima di un particolare obiettivo.

Sig. Valore di significatività della statistica di Ljung-Box. Un valore di significatività inferiore a 0.05 indica che gli errori residui non sono casuali.

Statistiche di riepilogo. Questa sezione contiene una serie di statistiche di riepilogo per le varie colonne, compresi i valori medi, minimi, massimi e percentili.

Parametri del modello di serie temporali

La scheda Parametri elenca i dettagli dei vari parametri utilizzati per creare un modello selezionato.

Visualizza parametri per il modello. Selezionare il modello per il quale si desidera visualizzare i dettagli dei parametri.

Obiettivo. Il nome del campo obiettivo (con ruolo *Obiettivo*) previsto da questo modello.

Modello. Il tipo di modello utilizzato per il campo obiettivo.

Campo (solo modelli ARIMA). Contiene una voce per ciascuna delle variabili utilizzate nel modello, con l'obiettivo per primo seguito dagli eventuali predittori.

Trasformazione. Indica quale tipo di trasformazione è stato specificato per il campo prima della creazione del modello.

Parametro. Il parametro del modello per il quale sono visualizzati i seguenti dettagli:

- **Ritardo (solo modelli ARIMA).** Indica gli eventuali ritardi considerati per questo parametro nel modello.
- **Stima.** La stima dei parametri. Questo valore è utilizzato nel calcolo del valore di previsione e degli intervalli di confidenza per il campo obiettivo.
- **SE.** L'errore standard della stima dei parametri.
- **t.** Il valore della stima del parametro diviso per l'errore standard.
- **Sig.** Il livello di significatività della stima dei parametri. I valori superiori a 0.05 sono considerati non statisticamente significativi.

Residui dei modelli di serie temporali

La scheda Residui mostra le funzioni di autocorrelazione (ACF) e di autocorrelazione parziale (PACF) dei residui (differenze tra valori previsti e valori effettivi) per ogni modello creato. Per ulteriori informazioni, consultare l'argomento "Funzioni di autocorrelazione e autocorrelazione parziale" a pagina 260.

Visualizza plot del modello. Selezionare il modello per cui si desidera visualizzare l'ACF e la PACF dei residui.

Riepilogo del modello di serie temporali

La scheda Riepilogo di un nugget del modello visualizza le informazioni sul modello stesso (*Analisi*), i campi utilizzati nel modello (*Campi*), le impostazioni utilizzate durante la creazione del modello (*Impostazioni di creazione*) e l'addestramento del modello (*Riepilogo addestramento*).

Quando si sfoglia per la prima volta il nodo, i risultati della scheda Riepilogo sono compressi. Per visualizzare i risultati, utilizzare il controllo dell'espansore a sinistra di un elemento se si desidera visualizzare solo i risultati di tale elemento oppure fare clic sul pulsante **Espandi tutto** se si desidera visualizzare tutti i risultati. Per nascondere i risultati una volta terminata la visualizzazione, utilizzare il controllo dell'espansore per comprimere i risultati specifici che si desidera nascondere oppure fare clic sul pulsante **Comprimi tutto** per comprimere tutti i risultati.

Analisi. Visualizza informazioni sul modello specifico.

Campi. Elenca i campi utilizzati come obiettivi e come input nella creazione del modello.

Impostazioni di creazione. Contiene informazioni sulle impostazioni utilizzate nella creazione del modello.

Riepilogo addestramento. Mostra il tipo di modello, il flusso utilizzato per creare tale modello, l'utente che lo ha creato, la data di creazione e il tempo trascorso per la creazione del modello.

Impostazioni del modello di serie temporali

La scheda Impostazioni consente di specificare i campi aggiuntivi generati dall'operazione di creazione del modello.

Crea nuovi campi per ciascun modello di cui calcolare il punteggio. Consente di specificare i nuovi campi da creare per ogni modello di cui calcolare il punteggio.

- **Calcola i limiti di confidenza superiore e inferiore.** Se è selezionata, questa opzione crea dei nuovi campi (con i prefissi di default $TSLCI- e $TSUCI-) per gli intervalli di confidenza inferiore e superiore, rispettivamente, per ogni campo obiettivo, insieme ai totali di detti valori.
- **Calcola residui di rumore.** Se è selezionata, questa opzione crea un nuovo campo (con il prefisso di default $TSNR-) per i residui del modello di ogni campo obiettivo, insieme al totale di detti valori.

Capitolo 14. Modelli nodo Risposta autoapprendimento

Nodo SLRM

Il nodo **Modello risposta autoapprendimento** (SLRM) consente di creare un modello che è possibile aggiornare continuamente o stimare nuovamente con l'aumentare delle dimensioni dell'insieme di dati senza dovere ricreare ogni volta il modello utilizzando l'insieme di dati completo. Per esempio, questo modello è utile quando si dispone di numerosi prodotti e si desidera identificare quale prodotto è più probabile che un cliente acquisti qualora gli venga offerto. Questo modello consente di prevedere quali sono le offerte più adatte per i clienti e qual è la probabilità che vengano accettate.

All'inizio, il modello può essere creato a partire da un insieme di dati ristretto con una serie di offerte proposte a caso e le risposte a tali offerte. Con l'aumentare delle dimensioni dell'insieme di dati, il modello può essere aggiornato e quindi migliora la sua capacità di prevedere quali sono le offerte più adatte per i clienti e qual è la probabilità che vengano accettate in base ad altri campi di input quali età, sesso, occupazione e reddito. Le offerte disponibili si possono cambiare aggiungendole o rimuovendole dalla finestra di dialogo del nodo, senza che sia necessario modificare il campo obiettivo dell'insieme di dati.

Se abbinati a IBM SPSS Collaboration and Deployment Services, consentono di impostare aggiornamenti periodici automatici del modello. Questo processo, che non richiede la supervisione né l'intervento dell'utente, costituisce una soluzione flessibile ed economica per le organizzazioni e le applicazioni in cui l'intervento personalizzato di un data miner non è possibile o non è necessario.

Esempio. Una società finanziaria desidera ottenere risultati più redditizi nelle campagne future, inviando offerte mirate ai singoli clienti. È possibile utilizzare un modello di autoapprendimento per identificare le caratteristiche dei clienti con maggiori probabilità di rispondere in modo favorevole in base alle promozioni precedenti e per aggiornare il modello in tempo reale in base alle risposte dei clienti più recenti.

Opzioni dei campi del nodo SLRM

Prima di eseguire un nodo SLRM è necessario specificare sia il campo obiettivo che il campo risposta obiettivo nella scheda Campi del nodo.

Campo Obiettivo. Selezionare l'obiettivo dall'elenco, per esempio un campo nominale (insieme) contenente i vari prodotti che si desidera proporre ai clienti.

Nota: il campo obiettivo deve avere un'archiviazione di tipo stringa e non numerica.

Campo risposta obiettivo. Selezionare il campo risposta obiettivo dall'elenco. Per esempio, Accettato o Respinto.

Nota: questo campo deve essere di tipo Flag. Il valore vero del flag indica l'accettazione dell'offerta, mentre quello falso ne indica il rifiuto.

Gli altri campi di questa finestra di dialogo sono i campi normalmente utilizzati in IBM SPSS Modeler. Per ulteriori informazioni, consultare l'argomento “Opzioni dei campi dei nodi Modelli” a pagina 31.

Nota: se i dati di origine includono intervalli che devono essere utilizzati come campi di input (intervallo numerico) continui, è necessario verificare che i metadati includano i dettagli minimo e massimo per ciascun intervallo.

Opzioni del modello di nodo SLRM

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.

Utilizza dati partizionati. Se è definito un campo partizione, questa opzione garantisce che per la creazione del modello vengano utilizzati dati solo dalla partizione di addestramento.

Addestramento continuo modello esistente. Per default, a ogni esecuzione di un nodo Modelli viene creato un modello completamente nuovo. Se si seleziona questa opzione, l'addestramento continuerà con l'ultimo modello creato correttamente dal nodo. In questo modo sarà possibile aggiornare un modello esistente senza dover accedere ai dati originali, aumentando così significativamente le prestazioni poiché nel flusso verranno utilizzati *solo* i record nuovi o aggiornati. I dettagli del modello precedente vengono archiviati con il nodo Modelli, consentendo di utilizzare questa opzione anche se il nugget del modello precedente non è più disponibile nel flusso o nella palette Modelli.

Valori campo obiettivo. Per default, questa opzione è impostata su **Utilizza tutto**, per indicare che sarà creato un modello contenente ogni offerta associata al valore campo obiettivo selezionato. Se si desidera generare un modello contenente solo alcune delle offerte del campo obiettivo, fare clic su **Specifica** e utilizzare i pulsanti **Aggiungi**, **Modifica** ed **Elimina** per inserire o modificare i nomi delle offerte per cui si desidera creare un modello. Per esempio, se si sceglie un obiettivo che elenca tutti i prodotti forniti, è possibile utilizzare questo campo per limitare i prodotti offerti solo ai pochi prodotti immessi qui.

Valutazione del modello. I campi di questo riquadro sono indipendenti dal modello poiché non influiscono sul calcolo del punteggio. Essi consentono invece di creare una rappresentazione visiva della capacità del modello di prevedere i risultati.

Nota: per visualizzare i risultati della valutazione del modello nel nugget del modello occorre selezionare anche la casella **Visualizza valutazione del modello**.

- **Includi valutazione del modello.** Selezionare questa casella per creare grafici che mostrino la precisione prevista del modello per ogni offerta selezionata.
- **Imposta seed random.** Quando si esegue la stima della precisione di un modello 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. Inserire il valore seme desiderato. Se questa opzione non è selezionata, verrà generato un campione diverso ogni volta che si esegue il nodo.
- **Dimensioni campione simulato.** Specificare il numero di record da utilizzare nel campione durante la valutazione del modello. Il valore predefinito è 100.
- **Numero di iterazioni.** Questa opzione consente di interrompere la valutazione del modello dopo il numero di iterazioni specificate. Specificare il numero massimo di iterazioni; il valore di default è 20.

  Nota: tenere presente che campioni di grandi dimensioni e un numero elevato di iterazioni aumenteranno il tempo impiegato per creare il modello.

Visualizza valutazione del modello. Selezionare questa opzione per visualizzare una rappresentazione grafica dei risultati del nugget del modello.

Opzioni di impostazione del nodo SLRM

Le opzioni di impostazione del nodo consentono di definire con precisione il processo di creazione dei modelli.

Numero massimo di previsioni per record. Questa opzione consente di limitare il numero di previsioni eseguite per ciascun record dell'insieme di dati. Il valore predefinito è 3.

Si supponga per esempio di disporre di sei offerte (risparmio, prestito ipotecario, mutuo auto, pensione, carta di credito e assicurazione) e che si desideri conoscere le due migliori offerte da proporre ai clienti. In questo caso, si imposterà questo campo su 2. Quando il modello viene generato e associato a una tabella, vengono visualizzate due colonne di previsioni (insieme alla confidenza associata nella probabilità che l'offerta venga accettata) per ogni record. È possibile impostare fino a un massimo di sei previsioni per tutte e sei le possibili offerte.

Livello di randomizzazione. Per evitare eventuali distorsioni — ad esempio in un dataset incompleto o di piccole dimensioni — e considerare tutte le potenziali offerte in modo analogo, è possibile aggiungere un livello di randomizzazione alla selezione di offerte e alla probabilità di essere incluse come offerte consigliate. La randomizzazione è espressa in percentuale, come valore decimale compreso tra 0.0 (nessuna randomizzazione) e 1.0 (completamente casuale). Il valore di default è 0.0.

Imposta seed random. Quando si aggiunge un livello di randomizzazione alla selezione di un'offerta, 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. Inserire il valore seme desiderato. 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.

Criterio di ordinamento. Selezionare l'ordine in cui le offerte dovranno essere visualizzate nel modello generato:

- **Decrescente.** Il modello visualizza prima le offerte con i punteggi più alti. Si tratta delle offerte che hanno la maggiore probabilità di essere accettate.
- **Crescente.** Il modello visualizza prima le offerte con i punteggi più bassi. Si tratta delle offerte che hanno la maggiore probabilità di essere rifiutate. Per esempio, questa funzionalità potrebbe essere utile per decidere quali clienti rimuovere da una campagna di marketing per un'offerta specifica.

Preferenze per i campi obiettivo. Quando si crea un modello, è possibile che vi siano certi aspetti che si desidera promuovere oppure rimuovere. Per esempio, se si genera un modello che seleziona la migliore offerta finanziaria da pubblicizzare presso la clientela, è possibile garantire che una determinata offerta sia sempre inclusa indipendentemente dal punteggio che viene calcolato per ogni cliente.

Per includere un'offerta in questo riquadro e modificarne le preferenze, fare clic su **Aggiungi**, digitare il nome dell'offerta (per esempio, Risparmio o Prestito ipotecario) e fare clic su **OK**.

- **Valore.** Viene mostrato il nome dell'offerta aggiunta.
- **Preferenza.** Specificare il livello di preferenza da applicare all'offerta. La preferenza è espressa in percentuale, come valore decimale compreso tra 0.0 (nessuna preferenza) e 1.0 (massima preferenza). Il valore di default è 0.0.
- **Includi sempre.** Per garantire che una determinata offerta sia sempre inclusa nelle previsioni, selezionare questa casella.

  Nota: se **Preferenza** è impostata su 0.0, l'impostazione **Includi sempre** viene ignorata.

Considera affidabilità del modello. Un modello ricco di dati e ben strutturato che è stato perfezionato attraverso molteplici rigenerazioni produrrà sempre risultati più accurati rispetto a un modello nuovo con pochi dati. Per sfruttare la maggiore affidabilità del modello ottimizzato, selezionare questa casella.

Nugget del modello SLRM

Nota: in questa scheda, i risultati sono visualizzati solo se si seleziona sia **Includi valutazione del modello** che **Visualizza valutazione del modello** nella scheda delle opzioni del modello.

Quando si esegue un flusso contenente un modello SLRM, il nodo stima la precisione delle previsioni per il valore di ogni campo obiettivo (offerta) e l'importanza di ogni predittore utilizzato.

Nota: se nella scheda Modello del nodo Modelli è stata selezionata l'opzione **Addestramento continuo modello esistente**, le informazioni visualizzate nel nugget del modello vengono aggiornate ogni volta che si rigenera il modello.

Per i modelli creati con IBM SPSS Modeler versione 12.0 o superiore, la scheda Modello del nugget del modello è suddivisa in due colonne:

Colonna sinistra.

- **Visualizza.** Quando si dispone di più offerte, selezionare quella di cui si desidera visualizzare i risultati.
- **Prestazione modello.** Mostra la precisione delle singole offerte stimata dal modello. L'insieme di test viene generato attraverso la simulazione.

Colonna destra.

- **Visualizza.** Selezionare se si desidera visualizzare i dettagli **Associazione a risposta** o **Importanza delle variabili**.
- **Associazione a risposta.** Mostra l'associazione (correlazione) dei singoli predittori con la variabile obiettivo.
- **Importanza predittore.** Indica l'importanza relativa dei singoli predittori nella stima del modello. In genere è opportuno concentrare la modellazione sui predittori più importanti e valutare se eliminare o ignorare quelli meno importanti. Questo grafico può essere interpretato nello stesso modo di altri modelli che mostrano l'importanza dei predittori, sebbene nel caso di SLRM il grafico venga generato per simulazione dall'algoritmo SLRM. L'impatto viene calcolato eliminando a turno ogni predittore dal modello e verificando in che modo questo influisce sulla precisione del modello stesso. Per ulteriori informazioni, consultare l'argomento "Importanza predittore" a pagina 43.

Impostazioni del modello SLRM

La scheda Impostazioni di un nugget del modello SLRM contiene le opzioni per la modifica del modello creato. Per esempio, è possibile utilizzare il nodo SLRM per creare vari modelli diversi a partire dagli stessi dati e dalle stesse impostazioni e quindi utilizzare questa scheda in ogni modello per modificare leggermente le impostazioni al fine di verificare in che modo questo influisce sui risultati.

Nota: questa scheda è disponibile solo dopo che il nugget del modello è stato aggiunto a un flusso.

Numero massimo di previsioni per record. Questa opzione consente di limitare il numero di previsioni eseguite per ciascun record dell'insieme di dati. Il valore predefinito è 3.

Si supponga per esempio di disporre di sei offerte (risparmio, prestito ipotecario, mutuo auto, pensione, carta di credito e assicurazione) e che si desideri conoscere le due migliori offerte da proporre ai clienti. In questo caso, si imposterà questo campo su 2. Quando il modello viene generato e associato a una tabella, vengono visualizzate due colonne di previsioni (insieme alla confidenza associata nella probabilità che l'offerta venga accettata) per ogni record. È possibile impostare fino a un massimo di sei previsioni per tutte e sei le possibili offerte.

Livello di randomizzazione. Per evitare eventuali distorsioni — ad esempio in un dataset incompleto o di piccole dimensioni — e considerare tutte le potenziali offerte in modo analogo, è possibile aggiungere

un livello di randomizzazione alla selezione di offerte e alla probabilità di essere incluse come offerte consigliate. La randomizzazione è espressa in percentuale, come valore decimale compreso tra 0.0 (nessuna randomizzazione) e 1.0 (completamente casuale). Il valore di default è 0.0.

Imposta seed random. Quando si aggiunge un livello di randomizzazione alla selezione di un'offerta, 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. Inserire il valore seme desiderato. 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.

Criterio di ordinamento. Selezionare l'ordine in cui le offerte dovranno essere visualizzate nel modello generato:

- **Decrescente.** Il modello visualizza prima le offerte con i punteggi più alti. Si tratta delle offerte che hanno la maggiore probabilità di essere accettate.
- **Crescente.** Il modello visualizza prima le offerte con i punteggi più bassi. Si tratta delle offerte che hanno la maggiore probabilità di essere rifiutate. Per esempio, questa funzionalità potrebbe essere utile per decidere quali clienti rimuovere da una campagna di marketing per un'offerta specifica.

Preferenze per i campi obiettivo. Quando si crea un modello, è possibile che vi siano certi aspetti che si desidera promuovere oppure rimuovere. Per esempio, se si genera un modello che seleziona la migliore offerta finanziaria da pubblicizzare presso la clientela, è possibile garantire che una determinata offerta sia sempre inclusa indipendentemente dal punteggio che viene calcolato per ogni cliente.

Per includere un'offerta in questo riquadro e modificarne le preferenze, fare clic su **Aggiungi**, digitare il nome dell'offerta (per esempio, Risparmio o Prestito ipotecario) e fare clic su **OK**.

- **Valore.** Viene mostrato il nome dell'offerta aggiunta.
- **Preferenza.** Specificare il livello di preferenza da applicare all'offerta. La preferenza è espressa in percentuale, come valore decimale compreso tra 0.0 (nessuna preferenza) e 1.0 (massima preferenza). Il valore di default è 0.0.
- **Includi sempre.** Per garantire che una determinata offerta sia sempre inclusa nelle previsioni, selezionare questa casella.

  Nota: se **Preferenza** è impostata su 0.0, l'impostazione **Includi sempre** viene ignorata.

Considera affidabilità del modello. Un modello ricco di dati e ben strutturato che è stato perfezionato attraverso molteplici rigenerazioni produrrà sempre risultati più accurati rispetto a un modello nuovo con pochi dati. Per sfruttare la maggiore affidabilità del modello ottimizzato, selezionare questa casella.

Capitolo 15. Modelli Support Vector Machine

Informazioni su SVM

Support Vector Machine (SVM) è una tecnica di classificazione e regressione robusta che ottimizza la precisione predittiva di un modello senza sovradattare i dati di addestramento. SVM si presta particolarmente all'analisi dei dati con numeri molto elevati di campi predittori (per esempio, migliaia).

SVM trova applicazione in molte discipline, incluse la gestione delle relazioni con i clienti (CRM), il riconoscimento facciale e di altre immagini, la bioinformatica, l'estrazione di concetti di text mining, il rilevamento delle intrusioni, la previsione della struttura delle proteine e il riconoscimento vocale.

Funzionamento di SVM

SVM mappa i dati a uno spazio di funzioni altamente dimensionale in modo che sia possibile categorizzare i punti dati, anche quando i dati non sono separabili linearmente in altro modo. Viene trovato un separatore tra le categorie e quindi i dati vengono trasformati in modo che il separatore possa essere tracciato come un iperpiano. In seguito, è possibile utilizzare le caratteristiche dei nuovi dati per prevedere il gruppo al quale deve appartenere un nuovo record.

Ad esempio, considerare la seguente figura, in cui i punti dati rientrano in due categorie differenti.

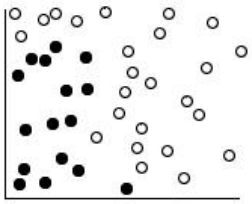


Figura 60. Insieme di dati originale

Le due categorie possono essere separate con una curva, come illustrato nella figura riportata di seguito.

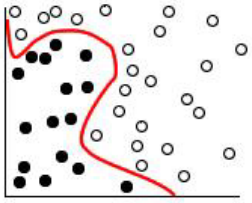


Figura 61. Dati con separatore aggiunto

Dopo la trasformazione, il limite tra le due categorie può essere definito mediante un iperpiano, come illustrato nella seguente figura.

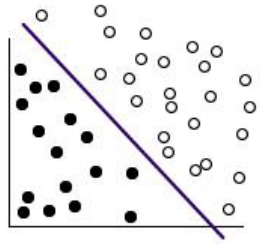


Figura 62. Dati trasformati

La funzione matematica utilizzata per la trasformazione è nota come funzione **Kernel**. In IBM SPSS Modeler, SVM supporta i seguenti tipi di Kernel:

- Lineare
- Polynomial
- RBF (Radial basis function)
- Sigmoid

Quando la separazione lineare dei dati è semplice, è consigliata una funzione Kernel lineare. Negli altri casi, è necessario utilizzare una delle altre funzioni. È necessario sperimentare le diverse funzioni per ottenere il miglior modello in ogni caso, poiché ogni funzione utilizza algoritmi e parametri diversi.

Ottimizzazione di un modello SVM

Oltre alla linea di separazione tra le categorie, un modello SVM di classificazione contiene anche linee marginali che definiscono lo spazio tra le due categorie.

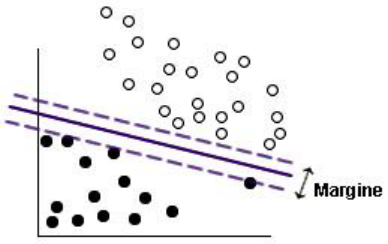


Figura 63. Dati con un modello preliminare

I punti dati che si trovano ai margini sono noti come **vettori di supporto**.

Maggiore è il margine tra le due categorie, più efficace sarà il modello nel prevedere la categoria dei nuovi record. Nell'esempio precedente, il margine non è molto ampio; si dice pertanto che il modello è **sovradattato**. Per ampliare il margine, è possibile accettare piccoli errori di classificazione; un esempio è riportato nella seguente figura.

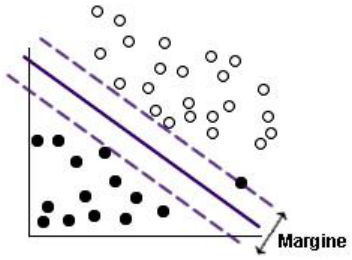


Figura 64. Dati con un modello migliorato

In alcuni casi, la separazione lineare è più difficile; un esempio è riportato nella seguente figura.

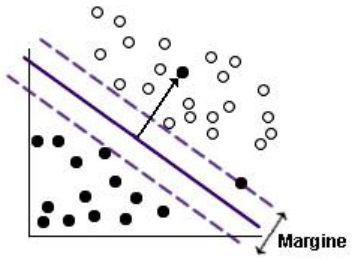


Figura 65. Problema di separazione lineare

In casi come questo, l'obiettivo è trovare l'equilibrio ottimale tra un margine ampio e un basso numero di punti dati classificati erroneamente. La funzione Kernel dispone di un **parametro di regolarizzazione** (noto come parametro C) che controlla il rapporto tra questi due valori. Probabilmente sarà necessario sperimentare valori diversi di questo e altri parametri Kernel per poter individuare il modello ottimale.

Nodo SVM

Il nodo SVM consente di utilizzare un algoritmo SVM per classificare i dati. SVM è particolarmente indicato per l'utilizzo con insiemi di dati di grandi dimensioni, cioè quelli con un elevato numero di campi predittore. È possibile utilizzare le impostazioni predefinite del nodo per produrre un modello di base in tempi relativamente rapidi, oppure utilizzare le impostazioni avanzate per sperimentare tipi diversi di modelli SVM.

Una volta creato il modello, è possibile:

- Sfogliare il nugget del modello per visualizzare l'importanza relativa dei campi di input nella creazione del modello.
- Accodare un nodo Tabella al nugget del modello per visualizzare l'output del modello.

Esempio. Un ricercatore medico ha ricevuto un insieme di dati contenente le caratteristiche di numerosi campioni di cellule umane estratte da pazienti ritenuti a rischio di sviluppo di tumori. L'analisi dei dati originali ha dimostrato che molte caratteristiche erano sostanzialmente diverse a seconda dei campioni benigni o maligni. Il ricercatore desidera sviluppare un modello SVM che può utilizzare i valori di caratteristiche cellulari simili nei campioni di altri pazienti per fornire un'indicazione precoce della benignità o malignità dei campioni.

Opzioni del modello di nodo SVM

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.

Utilizza dati partizionati. Se è definito un campo partizione, questa opzione garantisce che per la creazione del modello vengano utilizzati dati solo dalla partizione di addestramento.

Crea modelli di suddivisione. Crea un modello separato per ogni valore possibile dei campi di input che sono stati specificati come campi di suddivisione. Per ulteriori informazioni, consultare l'argomento "Creazione di modelli di suddivisione" a pagina 28.

Opzioni avanzate del nodo SVM

Le opzioni avanzate consentono agli utenti esperti di SVM di ottimizzare il processo di addestramento. Per accedere alle opzioni avanzate, impostare Moda su **Livello avanzato** nella scheda Livello avanzato.

Accoda tutte le probabilità (valido solo per gli obiettivi categoriali). Se questa opzione è selezionata (contrassegnata) specifica che le probabilità per ciascun valore possibile di un campo obiettivo flag o nominale vengono visualizzate per ciascun record elaborato dal nodo. Se questa opzione non è selezionata, per i campi obiettivo di tipo nominale o flag viene visualizzata soltanto la probabilità del valore previsto. L'impostazione di questa casella di controllo determina lo stato di default della corrispondente casella di controllo nella visualizzazione del nugget del modello.

Criteri di arresto. Determina quando interrompere l'algoritmo di ottimizzazione. I valori sono compresi tra 1.0E–1 e 1.0E–6; il valore predefinito è 1.0E–3. Riducendo il valore, si ottiene un modello più accurato, ma che richiede un addestramento più lungo.

Parametro di regolarizzazione (C). Controlla il rapporto tra l'ottimizzazione del margine e la riduzione del termine di errore di addestramento. Generalmente il valore deve essere compreso tra 1 e 10 inclusi; il valore predefinito è 10. Aumentando il valore si migliora il livello di precisione della classificazione (o si riduce l'errore di regressione) dei dati di addestramento, ma si può generare anche un sovradattamento.

Precisione di regressione (epsilon). Utilizzata solo se il livello di misurazione del campo obiettivo è *Continuo*. Determina l'accettazione degli errori, purché siano inferiori al valore qui specificato. L'aumento del valore può determinare una modellazione più rapida, ma a discapito della precisione.

Tipo Kernel. Determina il tipo di funzione Kernel utilizzato per la trasformazione. Tipi di Kernel diversi determinano modalità diverse di calcolo del separatore, pertanto è consigliabile sperimentare varie opzioni. Il valore di default è **RBF** (Radial Basis Function).

Gamma RBF. Attivata solo se il tipo di Kernel è **RBF**. In genere, il valore è compreso tra $3/k$ e $6/k$, dove k è il numero di campi di input. Per esempio, se vi sono 12 campi di input, è opportuno provare valori compresi tra 0.25 e 0.5. Aumentando il valore si migliora il livello di precisione della classificazione (o si riduce l'errore di regressione) dei dati di addestramento, ma si può generare anche un sovradattamento.

Gamma. Attivato solo se il tipo di Kernel è **Polinomiale** o **Sigmoide**. Aumentando il valore si migliora il livello di precisione della classificazione (o si riduce l'errore di regressione) dei dati di addestramento, ma si può generare anche un sovradattamento.

Distorsione. Attivato solo se il tipo di Kernel è **Polinomiale** o **Sigmoide**. Imposta il valore `coef0` nella funzione Kernel. Il valore di default 0 è adatto nella maggior parte dei casi.

Grado. Attivata solo se il tipo di Kernel è **Polinomiale**. Controlla la complessità (dimensione) dello spazio di mappatura. In genere non si utilizza un valore superiore a 10.

Nugget del modello SVM

Il modello SVM crea numerosi nuovi campi. Il più importante è il campo **$S-nomecampo**, che mostra il valore del campo obiettivo previsto dal modello.

Il numero e i nomi dei nuovi campi creati dal modello dipendono dal livello di misurazione del campo obiettivo (indicato nelle tabelle che seguono da *nomecampo*).

Per visualizzare questi campi e i loro valori, aggiungere un nodo Tabella al nugget del modello SVM ed eseguire il nodo Tabella.

Tabella 24. Il livello di misurazione del campo obiettivo è 'Nominale' o 'Flag'

Nome nuovo campo	Descrizione
$S-*nome campo*	Valore previsto del campo obiettivo.
$SP-*nome campo*	Probabilità del valore previsto.
$SP-*valore*	Probabilità di ogni possibile valore nominale o flag (visualizzato solo se è selezionata la casella di controllo **Accoda tutte le probabilità** nella scheda Impostazioni del nugget del modello).
$SRP-*valore*	(Solo obiettivi flag) Punteggi di propensione grezza (SRP) e regolata (SAP), che indicano la verosimiglianza di un risultato "vero" per il campo obiettivo. Questi punteggi vengono visualizzati solo se sono selezionate le caselle di controllo corrispondenti nella scheda Analisi del nodo Modello SVM prima della generazione del modello. Per ulteriori informazioni, consultare l'argomento “Opzioni della scheda Analizza di un nodo Modelli” a pagina 34.
$SAP-*valore*	

Tabella 25. Il livello di misurazione del campo obiettivo è 'Continuo'

Nome nuovo campo	Descrizione
$S-*nome campo*	Valore previsto del campo obiettivo.

Importanza predittore

Facoltativamente, nella scheda Modello, è possibile visualizzare anche un grafico che indica l'importanza relativa di ogni predittore nella stima del modello. In genere è opportuno concentrare la modellazione sui predittori più importanti e valutare se eliminare o ignorare quelli meno importanti. Notare che tale grafico è disponibile solo se è selezionata l'opzione **Calcola importanza predittore** nella scheda Analizza prima di generare il modello. Per ulteriori informazioni, consultare l'argomento “Importanza predittore” a pagina 43.

Nota: per calcolare l'importanza predittore per SVM potrebbe essere necessario più tempo rispetto a quanto necessario per il calcolo per altri tipi di modelli; inoltre, l'importanza predittore non è selezionata nella scheda Analizza per impostazione predefinita. La selezione di questa opzione potrebbe rallentare le prestazioni, in particolare con insiemi di dati di grandi dimensioni.

Impostazioni del modello SVM

La scheda Impostazioni consente di specificare campi aggiuntivi da visualizzare al momento della visualizzazione dei risultati (per esempio eseguendo un nodo Tabella collegato al nugget). Per visualizzare l'effetto prodotto da ciascuna di queste opzioni, selezionarle e fare clic sul pulsante Anteprima; scorrere l'anteprima dell'output verso destra per visualizzare i campi aggiuntivi.

Accoda tutte le probabilità (valido solo per gli obiettivi categoriali). Se è selezionata questa opzione, per ogni record elaborato dal nodo vengono visualizzate le probabilità di ogni possibile valore di un campo obiettivo di tipo nominale o flag. Se l'opzione è deselezionata, per i campi obiettivo di tipo nominale o flag vengono visualizzati soltanto il valore previsto e la sua probabilità.

L'impostazione di default di questa casella di controllo è determinata dalla corrispondente casella di controllo del nodo Modelli.

Calcola punteggi di propensione grezza. Per i modelli con obiettivi di tipo flag (che restituiscono un risultato sì o nessuna previsione), è possibile richiedere i punteggi di propensione grezza che indicano la verosimiglianza del risultato vero specificato per il campo obiettivo. Tali punteggi si aggiungono ai valori di previsione e confidenza che potrebbero venire generati durante il calcolo del punteggio.

Calcola punteggi di propensione regolata. I punteggi di propensione grezza sono basati esclusivamente sui dati di addestramento e potrebbero essere eccessivamente ottimistici a causa della tendenza di molti modelli a sovradattare questi dati. Le propensioni regolate cercano di operare una compensazione esaminando le prestazioni del modello rispetto a una partizione di test o di convalida. Per utilizzare questa opzione è necessario che nel flusso sia definito un campo partizione e che nel nodo Modelli siano attivati i punteggi di propensione regolati, prima di generare il modello.

Capitolo 16. Modelli dell'elemento adiacente più vicino

Nodo KNN

L'analisi della approssimità è un metodo che consente la classificazione dei casi in base alla loro somiglianza con altri casi. Questa analisi è stata sviluppata per l'apprendimento automatico, come metodo per riconoscere gli schemi di dati senza che sia necessaria una corrispondenza esatta con gli schemi, o i casi, archiviati. I casi simili sono vicini gli uni agli altri, mentre i casi dissimili sono distanti gli uni dagli altri. Pertanto, la distanza tra due casi è una misura della loro dissimilarità.

I casi vicini gli uni agli altri sono definiti "elementi adiacenti." Quando viene presentato un nuovo caso (holdout), viene calcolata la sua distanza da ognuno dei casi nel modello. Le classificazioni dei casi più simili – gli elementi adiacenti più vicini – vengono conteggiate ed il nuovo caso viene posizionato nella categoria che contiene il numero più alto di elementi adiacenti più vicini.

È possibile specificare il numero di elementi adiacenti più vicini da esaminare; tale valore è definito *k*. Le immagini mostrano il modo in cui viene classificato un nuovo caso utilizzando due valori differenti di *k*. Quando *k* = 5, il nuovo caso viene posizionato nella categoria *1* perché la maggioranza di elementi adiacenti più vicini appartiene alla categoria *1*. Tuttavia, quando *k* = 9, il nuovo caso viene posizionato nella categoria *0* perché la maggioranza di elementi adiacenti più vicini appartiene alla categoria *0*.

L'analisi della approssimità può anche essere usata per calcolare i valori per un target continuo. In questa situazione, per ottenere il valore previsto per il nuovo caso, viene utilizzato il valore obiettivo medio o mediano degli elementi adiacenti più vicini.

Opzioni degli obiettivi del nodo KNN

Nella scheda Obiettivi è possibile decidere se generare un modello che preveda il valore di un campo obiettivo nei dati di input in base ai valori degli elementi adiacenti più vicini, oppure se trovare semplicemente quali sono gli elementi adiacenti più vicini di un caso di interesse specifico.

Che tipo di analisi si desidera eseguire?

Prevedi un campo obiettivo. Scegliere questa opzione se si desidera prevedere il valore di un campo obiettivo in base ai valori degli elementi adiacenti più vicini.

Identifica solo gli elementi adiacenti più vicini. Scegliere questa opzione se si desidera visualizzare gli elementi adiacenti più vicini di un particolare campo di input.

Se si sceglie di identificare solo gli elementi adiacenti più vicini, le altre opzioni di questa scheda, quelle relative a precisione e velocità, vengono disabilitate in quanto sono rilevanti solo per la previsione degli obiettivi.

Qual è il proprio obiettivo?

Quando viene eseguita la previsione di un campo obiettivo, questo gruppo di opzioni consente di decidere se la velocità, l'accuratezza o una combinazione di entrambi i fattori rappresentano i fattori più importanti durante la previsione di un campo obiettivo. In alternativa, è possibile scegliere di personalizzare le impostazioni.

Se si seleziona Bilanciamento, Velocità o Precisione, l'algoritmo preseleziona la combinazione di impostazioni più indicata per quell'opzione. Gli utenti esperti possono, se lo desiderano, ignorare queste selezioni mediante i vari riquadri della scheda Impostazioni.

Bilancia velocità e accuratezza. Seleziona il numero ideale di elementi adiacenti in un intervallo ristretto.

Velocità. Trova un numero fisso di elementi adiacenti.

Accuratezza. Seleziona il numero ideale di elementi adiacenti in un intervallo più ampio e utilizza l'importanza dei predittori per il calcolo delle distanze.

Analisi personalizzata. Scegliere questa opzione per perfezionare l'algoritmo nella scheda Impostazioni.

Nota: la dimensione del modello KNN risultante, a differenza della maggior parte degli altri modelli, aumenta in modo lineare rispetto alla quantità di dati di addestramento. Se durante la creazione di un modello KNN viene visualizzato un errore di esaurimento memoria, provare a incrementare il valore della memoria di sistema massima utilizzata da IBM SPSS Modeler. A questo scopo, scegliere

Strumenti > **Opzioni** > **opzioni di sistema**

e immettere le nuove dimensioni nel campo **Memoria massima**. Le modifiche apportate in questa finestra di dialogo diventeranno effettive solo al successivo riavvio di IBM SPSS Modeler.

Impostazioni del nodo KNN

Nella scheda Impostazioni vengono selezionate le opzioni specifiche per l'analisi della approssimità. La barra laterale sulla sinistra dello schermo elenca i riquadri utilizzati per specificare le opzioni.

Modello

Il riquadro Modello fornisce le opzioni che controllano il modo in cui il modello deve essere creato; per esempio, la scelta se utilizzare il partizionamento o i modelli di suddivisione, se trasformare i campi di input numerici in modo che rientrino tutti nello stesso intervallo e la modalità di gestione dei casi di interesse. È possibile inoltre scegliere un nome personalizzato per il modello.

Nota: L'opzione **Utilizza dati partizionati** e **Utilizza etichette casi** non possono utilizzare lo stesso campo.

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.

Utilizza dati partizionati. Se è definito un campo partizione, questa opzione garantisce che per la creazione del modello vengano utilizzati dati solo dalla partizione di addestramento.

Crea modelli di suddivisione. Crea un modello separato per ogni valore possibile dei campi di input che sono stati specificati come campi di suddivisione. Per ulteriori informazioni, consultare l'argomento “Creazione di modelli di suddivisione” a pagina 28.

Per selezionare manualmente i campi... Per impostazione predefinita, il nodo utilizza le impostazioni relative alla partizione ed ai campi di suddivisione (se presenti) dal nodo Tipo, ma, in questo punto, è possibile sovrascrivere tali impostazioni. Per attivare i campi **Partizione** e **Suddivisioni**, selezionare la scheda **Campi** e scegliere **Usa impostazioni personalizzate**, quindi tornare in questo punto.

- **Partizione.** Questo campo consente di specificare un campo utilizzato per partizionare i dati in campioni distinti per le fasi di addestramento, di test e di convalida della creazione del modello. Utilizzando un campione per generare il modello e un altro campione per sottoporlo a verifica, è possibile ottenere un'indicazione valida del modo in cui il modello potrà essere esteso a insiemi di dati di dimensioni maggiori, simili ai dati correnti. Se sono stati definiti più campi partizione utilizzando nodi Tipo o Partizione, 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. Si noti inoltre che per applicare la partizione

selezionata nell'analisi, è necessario che il partizionamento sia attivato anche nella scheda Opzioni modello relativa al nodo. La disattivazione di questa opzione consente di disattivare il partizionamento senza dover modificare le impostazioni dei campi.

- **Suddivisioni.** Per i modelli di suddivisione, selezionare il campo o i campi di suddivisione. Questa operazione è simile all'impostazione del ruolo di un campo su *Suddivisione* in un nodo Tipo. È possibile designare solo i campi di tipo **Flag**, **Nominale** o **Ordinale** come campi di suddivisione. I campi selezionati come campi di suddivisione non possono essere utilizzati come campi obiettivo, di input, di partizione, di frequenza o peso. Per ulteriori informazioni, consultare l'argomento "Creazione di modelli di suddivisione" a pagina 28.

Normalizza input intervallo. Selezionare questa casella per normalizzare i valori per i campi di input continui. Le funzioni normalizzate hanno lo stesso intervallo di valori, il che può migliorare le prestazioni dell'algoritmo di stima. Viene utilizzata la normalizzazione corretta [2*(x−min)/(max−min)]−1. I valori della normalizzazione corretta sono compresi tra −1 e 1.

Utilizza etichette casi. Selezionare questa casella per abilitare l'elenco a discesa, da cui è possibile scegliere un campo i cui valori verranno utilizzati come etichette per identificare i casi di interesse nel grafico dello spazio predittori, nel grafico dei peer e nella mappa quadranti del visualizzatore del modello. È possibile designare come campi di etichettatura tutti i campi con livello di misurazione *Nominale*, *Ordinale* o *Flag*. Se non si sceglie un campo qui, i record vengono visualizzati nei grafici del visualizzatore del modello identificando gli elementi adiacenti in base al numero di riga nei dati di origine. Se dopo la creazione dei modelli i dati saranno ulteriormente manipolati, utilizzare le etichette dei casi per evitare ogni volta di dover consultare nuovamente i dati di origine per identificare i casi visualizzati.

Identifica record focale. Selezionare questa casella per abilitare l'elenco a discesa che consente di contrassegnare un campo di input di particolare interesse (solo per i campi flag). Se invece qui si specifica un campo, i punti che rappresentano il campo inizialmente saranno selezionati nel visualizzatore del modello quando viene creato il modello. La selezione di un record focale qui è facoltativa: qualsiasi punto può diventare temporaneamente un record focale se selezionato manualmente nel visualizzatore del modello.

Elementi adiacenti

Il riquadro Elementi adiacenti contiene una serie di opzioni che controllano il modo in cui viene calcolato il numero degli elementi adiacenti più vicini.

Numero di elementi adiacenti più vicini (k). Specifica il numero di elementi adiacenti più vicini relativamente a un caso specifico. L'utilizzo di un numero maggiore di elementi adiacenti non garantisce necessariamente un modello più preciso.

Se il proprio scopo è quello di prevedere un obiettivo, ci sono due possibilità:

- **Specifica k fisso.** Utilizzare questa opzione se si desidera specificare un numero fisso di elementi adiacenti più vicini da individuare.
- **Seleziona automaticamente k.** In alternativa, è possibile utilizzare i campi **Minimo** e **Massimo** per specificare un intervallo di valori e consentire alla procedura di scegliere il "miglior" numero di elementi adiacenti all'interno di tale intervallo. Il metodo per determinare il numero di elementi adiacenti dipende dall'eventualità che nel riquadro Selezione funzioni sia richiesta la selezione di funzioni:

  Se la selezione delle funzioni è attiva, viene eseguita per ciascun valore di *k* nell'intervallo richiesto e viene selezionato il *k* (con il relativo insieme di funzioni) con il tasso di errore più basso (o l'errore più basso della somma dei quadrati se l'obiettivo è continuo).

  Se la selezione delle funzioni non è attiva, viene utilizzata la convalida incrociata con occorrenze *V* per selezionare il "miglior" numero di elementi adiacenti. Vedere il riquadro Convalida incrociata per il controllo sull'assegnazione delle occorrenze.

Calcolo delle distanze. Metrica utilizzata per specificare la metrica di distanza per la misurazione della similarità dei casi.

- **Metrica euclidea.** La distanza tra due casi, x e y, è pari alla radice quadrata della somma, in tutte le dimensioni, dei quadrati delle differenze tra i valori di tali casi.
- **Metrica City Block.** La distanza tra due casi è pari alla somma, in tutte le dimensioni, delle differenze assolute tra i valori di tali casi. È denominata anche "distanza di Manhattan".

Facoltativamente, se il proprio scopo è quello di prevedere un obiettivo, è possibile scegliere di ponderare le funzioni in base alla loro importanza normalizzata nel calcolo delle distanze. L'importanza delle funzioni per un predittore è calcolata mediante il rapporto tra il tasso di errore o la somma dei quadrati degli errori del modello in cui il predittore sia stato eliminato e il tasso di errori o la somma dei quadrati degli errori del modello completo. L'importanza normalizzata si calcola riponderando i valori di importanza della funzione in modo che la somma sia pari a 1.

Pondera le caratteristiche in base all'importanza durante il calcolo delle distanze. (Visualizzato solo se si ha come scopo la previsione di un obiettivo.) Selezionare questa casella se si desidera che l'importanza dei predittori venga utilizzata nel calcolo delle distanze tra elementi adiacenti. L'importanza dei predittori verrà quindi visualizzata nel nugget del modello e utilizzata nelle previsioni (e influirà così sul punteggio). Per ulteriori informazioni, consultare l'argomento “Importanza predittore” a pagina 43.

Previsioni per obiettivo intervallo. (Visualizzato solo se si ha come scopo la previsione di un obiettivo.) Specificando un target continuo (intervallo numerico) si stabilisce se il valore previsto viene calcolato in base alla media o al valore mediano degli elementi adiacenti.

Selezione funzioni

Questo riquadro è visualizzato solo se lo scopo è quello di prevedere un obiettivo. Consente di richiedere e specificare opzioni per la selezione di funzioni. Per impostazione predefinita, per la selezione delle funzioni vengono prese in considerazione tutte le funzioni, ma è possibile selezionare un sottoinsieme di funzioni da forzare nel modello.

Effettua selezione delle funzioni. Selezionare questa casella per abilitare le opzioni di selezione delle funzioni.

- **Inserimento forzato.** Fare clic sul pulsante del selettore di campo accanto a questa casella e scegliere una o più funzioni da inserire forzatamente nel modello.

Criterio di arresto. Di volta in volta, viene presa in considerazione, per essere inclusa nell'insieme dei modelli, la funzione il cui inserimento nel modello determina l'errore minore (calcolato come tasso di errore per gli obiettivi categoriali e come errore della somma dei quadrati per i target continui). La selezione Forward prosegue fino al raggiungimento della condizione specificata.

- **Interrompe quando è stato selezionato il numero di caratteristiche specificato.** L'algoritmo inserisce un numero fisso di funzioni oltre a quelle forzate nel modello. Specificare un intero positivo. La riduzione dei valori del numero da selezionare dà origine a un modello più parsimonioso, con il rischio di perdere funzioni importanti. L'aumento dei valori del numero da selezionare consente di acquisire tutte le funzioni importanti, con il rischio però di aggiungere funzioni che finiscono per moltiplicare l'errore del modello.
- **Interrompe quando la variazione nel rapporto errore assoluto è inferiore o uguale al minimo.** L'algoritmo si arresta quando la variazione del rapporto di errore assoluto indica che il modello non può essere migliorato ulteriormente aggiungendo altre funzioni. Specificare un numero positivo. I valori decrescenti della variazione minima tendono a includere più funzioni con il rischio di includere funzioni che non aggiungono molto valore al modello. L'aumento del valore della variazione minima, invece, tende a impedire l'inserimento di altre funzioni, con il rischio di perderne alcune importanti per il modello. Il valore “ottimale” della modifica minima dipende dai dati e dall'applicazione. Per assistenza nella valutazione delle funzioni più importanti, vedere il registro degli errori relativi alla selezione delle funzioni nell'output. Per ulteriori informazioni, consultare l'argomento “Registro degli errori relativi alla selezione dei predittori” a pagina 295.

Convalida incrociata

Questo riquadro è visualizzato solo se lo scopo è quello di prevedere un obiettivo. Le opzioni presenti in questo riquadro consentono di decidere se utilizzare la convalida incrociata durante il calcolo degli elementi adiacenti più vicini.

La convalida incrociata divide il campione in varie **occorrenze**. I modelli di elementi adiacenti più vicini vengono quindi generati escludendo di volta in volta i dati da ciascun sottocampione. Il primo modello si basa su tutti i casi eccetto quelli contenuti nella prima occorrenza, il secondo modello si basa su tutti i casi eccetto quelli contenuti nella seconda occorrenza del campione e così via. Il rischio di errore per ciascun modello viene stimato applicando il modello al sottocampione escluso al momento della generazione del modello stesso. Il "miglior" numero di elementi adiacenti più vicini è quello che genera l'errore più basso in tutte le occorrenze.

Occorrenze con convalida incrociata. La convalida incrociata con occorrenze *V* viene utilizzata per determinare il "miglior" numero di elementi adiacenti. Per motivi legati alle prestazioni, la convalida incrociata non è disponibile se si utilizza la selezione delle funzioni.

- **Assegna in modo casuale i casi alle occorrenze.** Specificare il numero di occorrenze da utilizzare per la convalida incrociata. I casi vengono assegnati in modo casuale alle occorrenze, numerati da 1 a *V*, il numero di occorrenze.
- **Imposta seed random.** Quando si esegue la stima della precisione di un modello 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. Inserire il valore seme desiderato. Se questa opzione non è selezionata, verrà generato un campione diverso ogni volta che si esegue il nodo.
- **Utilizza campo per assegnare casi.** Specificare un campo numerico che assegni ad un'occorrenza ciascun caso dell'insieme di dati attivo. Il campo deve essere numerico e contenere valori compresi tra 1 e *V*. Se in questo intervallo mancano dei valori (o delle suddivisioni, se sono attivi i modelli di suddivisione) verrà generato un errore.

Analizza

Il riquadro Analizza è attivo solo se lo scopo è quello di prevedere un obiettivo. Questo riquadro può essere utilizzato per indicare se il modello dovrà includere altre variabili contenenti:

- le probabilità per ogni valore di campo obiettivo possibile
- le distanze tra un caso e gli elementi adiacenti più vicini
- i punteggi di propensione grezza e regolata (solo obiettivi flag)

Accoda tutte le probabilità. Se è selezionata questa opzione, per ogni record elaborato dal nodo vengono visualizzate le probabilità di ogni possibile valore di un campo obiettivo di tipo nominale o flag. Se l'opzione è deselezionata, per i campi obiettivo di tipo nominale o flag vengono visualizzati soltanto il valore previsto e la sua probabilità.

Salva distanze tra casi e gli elementi adiacenti più vicini k. Per ciascun record focale, viene creata una variabile per ciascuno degli elementi adiacenti più vicini *k* del record focale (dal campione di addestramento) e le distanze più vicine *k* corrispondenti.

Punteggi di propensione

I punteggi di propensione possono essere attivati nel nodo Modelli e nella scheda Impostazioni del nugget del modello. Questa funzionalità è disponibile solamente quando l'obiettivo selezionato è un campo flag. Per ulteriori informazioni, consultare l'argomento “Punteggi di propensione” a pagina 35.

Calcola punteggi di propensione grezza. I punteggi di propensione grezza vengono derivati dal modello in base ai soli dati di addestramento. Se il modello prevede il valore *vero* (risposta favorevole), la propensione sarà uguale a P, dove P è la probabilità della previsione. SE il modello prevede il valore falso, la propensione viene calcolata come (1 – P).

- Se si seleziona questa opzione durante la creazione del modello, nel nugget del modello i punteggi di propensione saranno attivati per default. Tuttavia, nel nugget del modello si può sempre decidere di attivare i punteggi di propensione grezza indipendentemente dal fatto che siano stati selezionati nel nodo Modelli.
- Quando si calcola il punteggio del modello, i punteggi di propensione grezza saranno aggiunti in un campo al cui prefisso standard vengono fatte seguire le lettere *RP*. Per esempio, se le previsioni sono in un campo denominato *$R-tasso di abbandono*, il nome del campo del punteggio di propensione sarà *$RRP-tasso di abbandono*.

Calcola punteggi di propensione regolata. Le propensioni grezze si basano unicamente sulle stime fornite dal modello, le quali possono essere sovradattate e determinare di conseguenza delle stime eccessivamente ottimistiche della propensione. Le propensioni regolate cercano di compensare esaminando le prestazioni del modello sulle partizioni di test e di convalida e adeguando le propensioni di conseguenza per fornire una stima più corretta.

- Questa impostazione richiede la presenza di un campo partizione valido nel flusso.
- Al contrario dei punteggi di confidenza grezza, i punteggi di propensione regolata devono essere calcolati in sede di creazione del modello, altrimenti non saranno disponibili quando si calcola il punteggio del nugget del modello.
- Quando si calcola il punteggio del modello, i punteggi di propensione regolata saranno aggiunti in un campo al cui prefisso standard vengono fatte seguire le lettere *AP*. Per esempio, se le previsioni sono in un campo denominato *$R-tasso di abbandono*, il nome del campo del punteggio di propensione sarà *$RAP-tasso di abbandono*. I punteggi di propensione regolata non sono disponibili per i modelli di regressione logistica.
- Quando si calcolano i punteggi di propensione regolata, la partizione di test o di convalida utilizzata per il calcolo non deve essere stata bilanciata. A tal fine, verificare che l'opzione **Bilancia solo dati di addestramento** sia selezionata in tutti i nodi bilanciamento a monte. Inoltre, se a monte è presente un campione complesso, i punteggi di propensione regolata saranno invalidati.
- I punteggi di propensione regolata non sono disponibili per i modelli di struttura ad albero boosted e di insiemi di regole. Per ulteriori informazioni, consultare l'argomento “Modelli C5.0 boosted” a pagina 113.

Nugget del modello KNN

Il modello KNN crea una serie di nuovi campi, come illustrato nella tabella seguente. Per visualizzare questi campi e i loro valori, aggiungere un nodo Tabella al nugget del modello KNN ed eseguire il nodo Tabella oppure fare clic sul pulsante Anteprima dell'insieme.

Tabella 26. Campi di modelli KNN

Nome nuovo campo	Descrizione
$KNN-*nome campo*	Valore previsto del campo obiettivo.
$KNNP-*nome campo*	Probabilità del valore previsto.
$KNNP-*valore*	Probabilità di ogni possibile valore di campo nominale o flag. Incluso solo se è selezionata la casella di controllo **Accoda tutte le probabilità** nella scheda Impostazioni del nugget del modello.
$KNN-neighbor-*n*	Il nome dell'*n* elemento adiacente più vicino rispetto al record focale. Incluso solo se, nella scheda Impostazioni del nugget del modello, l'opzione **Visualizza il più vicino** è stata impostata su un valore diverso da zero.

Tabella 26. Campi di modelli KNN (Continua)

Nome nuovo campo	Descrizione
$KNN-distance-*n*	La distanza relativa dal record focale dell'elemento adiacente più vicino *n* al record focale. Incluso solo se, nella scheda Impostazioni del nugget del modello, l'opzione **Visualizza il più vicino** è stata impostata su un valore diverso da zero.

Vista del modello dell'elemento adiacente più vicino

Vista Modello

con una finestra a due riquadri:

- Nel primo è presente una panoramica del modello denominata "vista principale".
- Nel secondo, invece, possono essere visualizzate due tipologie di vista:

  La vista ausiliaria mostra ulteriori informazioni sul modello, pur non concentrandosi su quest'ultimo.

  La vista collegata mostra invece i dettagli relativi a una funzione del modello quando l'utente esegue il drill-down di parte della vista principale.

Per default, nel primo riquadro viene visualizzato lo spazio dei predittori e nel secondo il grafico dell'importanza dei predittori. Se il grafico Importanza predittore non è disponibile, vale a dire quando **Pesa funzioni in base all'importanza** non è stata selezionata nel pannello Elementi adiacenti della scheda Impostazioni, viene mostrata la prima visualizzazione disponibile nel menu a discesa Visualizza.

Se per una vista non sono disponibili informazioni, la vista non appare nell'elenco a discesa.

Spazio dei predittori: Il grafico dello spazio dei predittori è un grafico interattivo relativo allo spazio dei predittori (o al sottospazio, se sono presenti più di tre predittori). Ogni asse rappresenta un predittore nel modello e la posizione dei punti nel grafico indica i valori di tali predittori per i casi nelle partizioni di addestramento e di holdout.

Chiavi. Oltre a rappresentare i valori dei predittori, i punti forniscono altre informazioni.

- La forma indica la partizione (Addestramento o Holdout) di cui fa parte un punto.
- Il colore/l'ombreggiatura di un punto indica il valore dell'obiettivo del caso (i diversi valori di colore corrispondono alle categorie di un obiettivo categoriale, mentre le ombreggiature indicano l'intervallo di valori di un target continuo). Il valore indicato per la partizione di addestramento è quello osservato, mentre per la partizione di holdout è indicato quello previsto. Se non viene specificato alcun obiettivo, questo simbolo non viene visualizzato.
- Contorni più marcati indicano che un caso è focale. I record focali vengono visualizzati collegati ai relativi elementi adiacenti più vicini *k*.

Comandi e interattività. Nel grafico è disponibile una serie di comandi per esplorare lo spazio dei predittori.

- È possibile scegliere il sottoinsieme di predittori da visualizzare nel grafico e cambiare i predittori da rappresentare nelle dimensioni.
- I “record focali” sono semplicemente punti selezionati nel grafico Spazio dei predittori. Se è stata specificata una variabile per record focali, inizialmente verranno selezionati i punti che rappresentano i record focali. Qualsiasi punto può comunque diventare temporaneamente un record focale se viene selezionato. Vengono utilizzati i controlli “usuali” per la selezione dei punti; facendo clic su un punto, tale punto viene selezionato e viene annullata la selezione di tutti gli altri punti; facendo clic su un punto tenendo premuto il tasto control, il punto viene aggiunto all'insieme dei punti selezionati. Le viste collegate, ad esempio il grafico dei peer, vengono automaticamente aggiornate in base ai casi selezionati nello spazio dei predittori.
- È possibile modificare il numero di elementi adiacenti più vicini(*k*) da visualizzare per i record focali.

- Passando il mouse sopra un punto del grafico, viene visualizzata una descrizione con il valore dell'etichetta del caso (o il numero del caso se non sono state definite etichette), oltre ai valori osservati e previsti dell'obiettivo.
- Un pulsante "Ripristina" consente di ripristinare lo stato originale dello spazio dei predittori.

Modifica degli assi nel grafico dello spazio dei predittori: È possibile decidere quali funzioni visualizzare sugli assi del grafico dello spazio dei predittori.

Per modificare le impostazioni degli assi:

1. Fare clic sul pulsante Modalità Modifica (l'icona a forma di pennello) nel riquadro di sinistra per selezionare la modalità Modifica per lo spazio dei predittori.
2. Modificare la visualizzazione (nel modo desiderato) nel riquadro di destra. Fra i due riquadri principali viene visualizzato il riquadro **Mostra zone**.
3. Selezionare la casella di controllo **Mostra zone**.
4. Fare clic su un punto dati qualsiasi nello spazio dei predittori.
5. Per sostituire un asse con un predittore dello stesso tipo di dati:
   - Trascinare il nuovo predittore sull'etichetta della zona (contrassegnata dal pulsantino con la X) del predittore che si desidera sostituire.
6. Per sostituire un asse con un predittore di un tipo di dati diverso:
   - Sull'etichetta della zona del predittore da sostituire, fare clic sul pulsantino con la X. La visualizzazione dello spazio dei predittori diventa bidimensionale.
   - Trascinare il nuovo predittore sull'etichetta della zona **Aggiungi dimensione**.
7. Fare clic sul pulsante della modalità Esplora (l'icona con la freccia) nel riquadro di sinistra per uscire dalla modalità Modifica.

Importanza predittore: Generalmente, lo sforzo della modellazione viene concentrato sui campi predittore più importanti, senza considerare o ignorando i campi di minore importanza. Il grafico dell'importanza dei predittori rende più semplice questa operazione, indicando l'importanza relativa di ciascun predittore nella stima del modello. Poiché i valori sono relativi, la somma dei valori visualizzata per tutti i predittori è 1.0. L'importanza dei predittori non è correlata alla precisione del modello. Riguarda unicamente l'importanza di ciascun predittore nell'esecuzione di una previsione e non il grado di precisione della previsione.

Distanze degli elementi adiacenti più vicini: In questa tabella vengono visualizzati gli elementi adiacenti più vicini e le distanze *k* solo per i record focali. La tabella è disponibile se un identificatore del record focale è stato specificato nel nodo Modelli e visualizza solo i record focali identificati da questa variabile.

Ogni riga della:

- Colonna **Record focale** contiene il valore della variabile di etichetta relativa al record focale. Se non sono definite etichette dei casi, la colonna contiene il numero di caso del record focale.
- La i^a colonna nel gruppo **Elementi adiacenti più vicini** contiene il valore della variabile etichetta del caso per il $i°$ elemento adiacente più vicino del record focale; se non sono state definite le etichette del caso, questa colonna contiene il numero del caso del $i°$ elemento adiacente più vicino del record focale.
- La i^a colonna nel gruppo **Distanze più prossime** contiene la distanza del $i°$ elemento adiacente più vicino al record focale.

Peer: In questo grafico vengono visualizzati i casi focali e i relativi elementi adiacenti più vicini *k* per ciascun predittore e per l'obiettivo. È disponibile se nello spazio dei predittori è selezionato un caso focale.

Il Grafico dei peer è collegato allo spazio dei predittori in due modi.

- I casi selezionati (focali) nello spazio dei predittori vengono visualizzati nel grafico dei peer, insieme ai relativi elementi adiacenti più vicini *k*.
- Il valore di *k* selezionato nello spazio dei predittori viene utilizzato nel grafico dei peer.

Seleziona predittori. Consente di selezionare i predittori da visualizzare nel Grafico dei peer.

Mappa dei quadranti: Il grafico mostra i casi focali e i relativi elementi adiacenti più vicini *k* su un grafico a dispersione (o un grafico a punti a seconda del livello di misurazione dell'obiettivo) con l'obiettivo sull'asse *y* e un predittore di scala sull'asse *x*, il tutto suddiviso in riquadri in base ai predittori. È disponibile se nello spazio dei predittori è presente un obiettivo ed è selezionato un caso focale.

- Per le variabili continue, nella partizione di addestramento in corrispondenza delle medie delle variabili vengono tracciate linee di riferimento.

Seleziona predittori. Consente di selezionare i predittori da visualizzare nella Mappa dei quadranti.

Registro degli errori relativi alla selezione dei predittori: I punti presenti nel grafico mostrano l'errore (in termini di tasso di errore o di errore della somma dei quadrati a seconda del livello di misurazione dell'obiettivo) sull'asse *y* del modello, con il predittore indicato sull'asse *x* (inoltre, a sinistra sull'asse *x* sono presenti tutte le funzioni). Il grafico è disponibile se è presente un obiettivo ed è attiva la selezione funzioni.

Tabella di classificazione: Nella tabella viene visualizzata la classificazione incrociata dei valori osservati dell'obiettivo rispetto a quelli previsti, suddivisi per partizione. È disponibile se esiste un obiettivo ed è categoriale (flag, nominale o ordinale).

- La riga **(Mancante)** della partizione di holdout contiene casi di controllo con valori mancanti sull'obiettivo. Tali casi contribuiscono ai valori di Campione di holdout: Percentuale globale ma non ai valori di Percentuale corretta.

Riepilogo degli errori: La tabella è disponibile in presenza di una variabile di destinazione. Visualizza l'errore associato al modello; somma dei quadrati per un target continuo e tasso di errore (100% – percentuale globale corretta) per un target di categoria.

Impostazioni del modello KNN

La scheda Impostazioni consente di specificare campi aggiuntivi da visualizzare al momento della visualizzazione dei risultati (per esempio eseguendo un nodo Tabella collegato al nugget). Per visualizzare l'effetto prodotto da ciascuna di queste opzioni, selezionarle e fare clic sul pulsante Anteprima; scorrere l'anteprima dell'output verso destra per visualizzare i campi aggiuntivi.

Accoda tutte le probabilità (valido solo per gli obiettivi categoriali). Se è selezionata questa opzione, per ogni record elaborato dal nodo vengono visualizzate le probabilità di ogni possibile valore di un campo obiettivo di tipo nominale o flag. Se l'opzione è deselezionata, per i campi obiettivo di tipo nominale o flag vengono visualizzati soltanto il valore previsto e la sua probabilità.

L'impostazione di default di questa casella di controllo è determinata dalla corrispondente casella di controllo del nodo Modelli.

Calcola punteggi di propensione grezza. Per i modelli con obiettivi di tipo flag (che restituiscono un risultato sì o nessuna previsione), è possibile richiedere i punteggi di propensione grezza che indicano la verosimiglianza del risultato vero specificato per il campo obiettivo. Tali punteggi si aggiungono ai valori di previsione e confidenza che potrebbero venire generati durante il calcolo del punteggio.

Calcola punteggi di propensione regolata. I punteggi di propensione grezza sono basati esclusivamente sui dati di addestramento e potrebbero essere eccessivamente ottimistici a causa della tendenza di molti modelli a sovradattare questi dati. Le propensioni regolate cercano di operare una compensazione esaminando le prestazioni del modello rispetto a una partizione di test o di convalida. Per utilizzare

questa opzione è necessario che nel flusso sia definito un campo partizione e che nel nodo Modelli siano attivati i punteggi di propensione regolati, prima di generare il modello.

Visualizza il più vicino. Se questo valore viene impostato su *n*, in cui *n* è un numero intero positivo e diverso da zero, gli *n* elementi adiacenti più vicini al record focale sono inclusi nel modello, assieme alle loro distanze relative dal record focale.

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 COMMERCIABILITÀ 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

A

AICC . Una misura per selezionare e confrontare modelli misti basata sulla verosimiglianza logaritmica -2 (ristretta). I valori più bassi indicano i modelli migliori. AICC corregge AIC in presenza di campioni piccoli. Mano mano che aumenta la dimensione del campione, il criterio AICC converge nel criterio AIC.

ANOVA univariate . Effettua l'analisi della varianza univariata a una via per verificare l'uguaglianza delle medie di gruppo per ciascuna variabile indipendente.

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.

B

BIC normalizzato . Criterio di informazione bayesiano normalizzato. Misura generale dell'adattamento generale di un modello che tenta di tenere conto della complessità del modello. Questa misura è un punteggio basato sull'errore quadratico medio e comprende una penalità per il numero di parametri del modello e una lunghezza per la serie. La penalità rimuove il vantaggio dei modelli con più parametri, consentendo di confrontare più facilmente la statistica con più modelli della stessa serie.

Bonferroni sequenziale . Una procedura di Bonferroni con scarti sequenzialmente decrescenti, molto meno conservativa in termini di rifiuto di singole ipotesi, ma che mantiene lo stesso livello di significatività globale.

C

Casi . Visualizza per ciascun caso i codici del gruppo effettivo, del gruppo previsto, della probabilità a posteriori e del punteggio discriminante.

Classificazione escludi uno . Ogni caso viene classificato usando le funzioni ricavate da tutti i casi meno se stesso. Nota anche come classificazione "metodo U".

Correlazione entro gruppi . Visualizza la matrice di correlazione entro gruppi ottenuta mediando le matrici di covarianza di tutti i gruppi prima di calcolare le correlazioni.

Covarianza . Una misura non standardizzata di associazione tra due variabili, pari alla deviazione del prodotto incrociato divisa per N-1.

Covarianza di gruppi separati . Visualizza matrici di covarianza separate per ciascun gruppo.

Covarianza entro gruppi . Visualizza una matrice combinata di covarianza entro i gruppi, che potrebbe differire dalla matrice di covarianza totale. La matrice è ottenuta dalla media delle matrici di covarianza di tutti i gruppi.

Covarianza totale . Visualizza una matrice di covarianza di tutti i casi come se provenissero da un unico campione.

Criterio informativo di Bayes (BIC) . Una misura per selezionare e confrontare modelli basata sulla verosimiglianza logaritmica -2. I valori più bassi indicano i modelli migliori. Anche BIC penalizza i modelli sovraparametrizzati, ma in modo più rigoroso rispetto ad AIC.

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.

D

Di Fisher . Visualizza i coefficienti di Fisher della funzione di classificazione, che possono essere usati direttamente per la classificazione. Viene riprodotto un insieme separato di coefficienti di funzioni di classificazione per ciascun gruppo. Ogni caso viene assegnato al gruppo in cui ottiene il più alto punteggio discriminante (valore della funzione di classificazione).

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.

Distanza di Mahalanobis . Una misura di quanto differiscano i valori di un caso per le variabili indipendenti, rispetto al valore medio di tutti i casi. Un'elevata distanza di Mahalanobis indica che un caso include valori estremi per una o più variabili indipendenti.

E

Entro i gruppi . La matrice di covarianza entro i gruppi viene utilizzato per classificare casi.

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 dell'assimetria . 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.

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).

G

Grafici di gruppi combinati . Crea un grafico a dispersione per tutti i gruppi dei primi due valori di funzioni discriminanti. Se esiste una sola funzione, viene invece visualizzato un istogramma.

Grafici di gruppi separati . Crea un grafico a dispersione per gruppi separati dei primi due valori di funzioni discriminanti. Se esiste una sola funzione, vengono invece visualizzati gli istogrammi.

Grafico di rischio . Visualizza la funzione di rischio cumulato in scala lineare.

Grafico di sopravvivenza . Visualizza la funzione di sopravvivenza cumulata in scala lineare.

Gruppi separati . Per classificare i casi vengono utilizzate le matrici di covarianza dei singoli gruppi. Dal momento che la classificazione è basata sulla funzione discriminante e non sui valori originali, questa opzione non è sempre equivalente alla discriminazione quadratica.

I

Intervallo . La differenza tra il valore massimo ed il valore minimo di una variabile numerica.

M

MAE . Errore assoluto della media. Misura di quanto la serie varia rispetto al livello predittivo del modello. Il valore MAE viene indicato in unità originali della serie.

MAPE . Errore percentuale assoluto medio. Indica di quanto la serie dipendente varia rispetto al livello predittivo del modello. L'errore è indipendente dalle unità usate e può quindi essere usato per confrontare le serie con unità diverse.

Mappa territoriale . Un grafico dei confini usati per classificare i casi in gruppi in base ai valori di una funzione. I numeri corrispondono ai gruppi nei quali vengono classificati i casi. La media per ciascun gruppo è indicata da un asterisco all'interno dei suoi confini. La mappa non viene visualizzata se c'è una sola funzione discriminante.

Massimizzazione del minimo rapporto F . Un metodo di selezione delle variabili nelle analisi per passi basato sulla massimizzazione di un rapporto F valutato tramite la distanza di Mahalanobis tra gruppi.

Massimo . Il valore più alto di una variabile numerica.

MaxAE . Errore assoluto massimo. L'errore stimato più alto, espresso nelle stesse unità della serie dipendente. Al pari di MaxAPE, è utile per stimare lo scenario peggiore per le previsioni. L'errore assoluto massimo e l'errore percentuale assoluto massimo possono verificarsi in punti diversi della serie; ad esempio se l'errore assoluto di una serie grande è leggermente maggiore dell'errore assoluto del valore minore della serie. In questo caso l'errore assoluto massimo si verifica nel valore maggiore della serie mentre l'errore percentuale assoluto massimo si verifica nel valore minore della serie.

MaxAPE . Errore percentuale assoluto massimo. Errore stimato più alto, espresso in percentuale. Questa misura è utile per stimare lo scenario peggiore delle previsioni.

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.

Medie . Visualizza le medie totali e di gruppo, nonché le deviazioni standard per le variabili indipendenti.

Minimizzazione del Lambda di Wilks . Un metodo di selezione delle variabili nell'analisi discriminante per passi che sceglie le variabili da inserire nell'equazione in base a quanto esse contribuiscono a minimizzare il Lambda di Wilks. Ad ogni passo viene inserita la variabile che minimizza il valore globale del Lambda di Wilks'.

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.

N

Non standardizzati . Visualizza i coefficienti della funzione discriminante non standardizzata.

R

R-quadrato . Misura della bontà dell'adattamento di un modello lineare, detto anche coefficiente di determinazione. È la proporzione di variabilità della variabile dipendente spiegata dal modello di regressione. Può variare tra 0 e 1. Valori bassi indicano che il modello non si adatta bene ai dati.

R-quadrato stazionario . Una misura che confronta la parte stazionaria del modello con un modello di media semplice. Questa misura è preferibile all'R-quadrato semplice se è presente una tendenza o un motivo stagionale. L'R-quadrato stazionario è negativo con un intervallo negativo infinito tendente a 1. I valori negativi indicano che il modello esaminato è peggiore del modello di base. I valori positivi indicano che il modello esaminato è migliore del modello di base.

Risultati di classificazione . Il numero di casi assegnati in modo corretto e non corretto a ciascuno dei gruppi in base all'analisi discriminante. A volte detta "Matrice confusione".

RMSE . Acronimo di Root Mean Square Error. La radice dell'errore quadratico medio. Indica di quanto una serie dipendente varia rispetto al livello stimato del modello. Questa misura viene espressa nelle stesse unità della serie dipendente.

S

Sidak sequenziale . Una procedura di Sidak con scarti sequenzialmente decrescenti, molto meno conservativa in termini di rifiuto di singole ipotesi, ma che mantiene lo stesso livello di significatività globale.

Somma . La somma o il totale di tutti i valori non mancanti di tutti i casi.

T

Test M di Box . Un test per l'uguaglianza di matrici di covarianza di gruppo. Per dimensioni di campione sufficientemente elevate, un valore P non significativo vuol dire che non ci sono sufficienti prove che le matrici differiscano. Il test è sensibile a scostamenti dalla normalità multivariata.

U

Varianza non spiegata . Ad ogni passo viene inserita la variabile che riduce al minimo la somma della variazione spiegata fra gruppi.

Univocità . Valuta tutti gli effetti simultaneamente, correggendo ogni effetto per tutti gli altri effetti di qualunque tipo.

Uno meno sopravvivenza . Visualizza il complemento a uno della funzione di sopravvivenza su una scala lineare.

Usa probabilità di F . La variabile viene inserita nel modello se il livello di significatività del relativo valore di F è minore di quello di inserimento. La variabile viene altresì rimossa se il livello di significatività è maggiore di quello di rimozione. I valori di inserimento e di rimozione devono essere entrambi positivi e Inserimento deve essere minore di Rimozione. Alzando il valore di inserimento e/o abbassando quello di rimozione Per rimuovere ulteriori variabili dal modello, riduce il valore di rimozione.

Utilizza valore F . La variabile viene inserita nel modello se il relativo valore F è maggiore di quello di inserimento. La variabile viene altresì rimossa se il relativo valore F è minore di quello di rimozione. I valori di inserimento e di rimozione devono essere entrambi positivi e Inserimento deve essere maggiore di Rimozione Abbassando il valore di inserimento e/o alzando quello di rimozione Per rimuovere ulteriori variabili dal modello, aumentare il valore di rimozione.

V

V di Rao (analisi discriminante) . Una misura delle differenze tra medie di gruppo. Detta anche traccia di Lawley-Hotelling. Ad ogni passo viene inserita la variabile che massimizza l'aumento della V di Rao. Dopo aver selezionato questa opzione, specificare l'incremento minimo che una variabile deve apportare per essere inserita nell'analisi.

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

E

F

G

I

K

L

M

N

O

P

R

S

T

V

Stampato in Italia