IBM SPSS Modeler 15 Guida alla modellazione in-database

Nota: Prima di utilizzare queste informazioni e il relativo prodotto, leggere le informazioni generali disponibili in Note a pag. .

Questa versione si applica a IBM SPSS Modeler 15 e a tutte le successive versioni e modifiche fino a eventuali disposizioni contrarie indicate in nuove versioni.

Le schermate dei prodotti Adobe sono state ristampate su autorizzazione di Adobe Systems Incorporated.

Le schermate dei prodotti Microsoft sono state ristampate su autorizzazione di Microsoft Corporation.

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 visiva di SPSS Modeler favorisce l'applicazione di una competenza aziendale specifica da parte degli utenti, grazie alla quale sarà possibile ottenere modelli di previsione più efficaci e una riduzione nei tempi di sviluppo delle soluzioni. SPSS Modeler offre una vasta gamma di tecniche di creazione di modelli, quali previsione, classificazione, segmentazione e 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 dell'azienda. 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 aziendali. 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 aziendali 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 contratto di manutenzione.

Contenuto

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 aziendali e di eseguirne il deployment nelle operazioni aziendali per migliorare i processi decisionali. Progettato secondo il modello CRISP-DM conforme agli standard di settore, SPSS Modeler supporta l'intero processo di data mining, dai dati a risultati aziendali migliori.

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, vedere*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
- IBM SPSS Modeler Server adattatori 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 trend, 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, vedere l'argomento Edizioni di IBM SPSS Modeler in *Manuale dell'utente di IBM SPSS Modeler 15*.

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 in-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 operazioni 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 uno stream SPSS Modeler che potrà essere eseguito da un motore di runtime esterno oppure incorporato in una applicazione esterna. Questo permette di pubblicare e sottoporre a deployment stream SPSS Modeler completi in ambienti in cui SPSS Modeler non è installato. SPSS Modeler Solution Publisher è distribuito come parte del servizio IBM SPSS Collaboration and Deployment

Services - Scoring, per cui è necessario procurarsi una licenza separata. Insieme alla licenza, si riceve SPSS Modeler Solution Publisher Runtime, che consente di eseguire gli stream pubblicati.

IBM SPSS Modeler Server Adattatori 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, uno stream 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 completamente nuova 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.

Text Analytics for IBM® SPSS® Modeler utilizza tecnologie linguistiche avanzate e di Natural Language Processing (NLP) per elaborare rapidamente una grande varietà di dati di testo non strutturati, estrarre e organizzare i concetti chiave e raggruppare questi 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 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=swg27023172*.

La documentazione in entrambi i formati è inoltre disponibile presso il Centro informazioni SPSS Modeler all'indirizzo *http://publib.boulder.ibm.com/infocenter/spssmodl/v15r0m0/*.

Documentazione di SPSS Modeler Professional

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

- **Manuale dell'utente di IBM SPSS Modeler.** Introduzione generale all'utilizzo di SPSS Modeler che illustra come creare stream di dati, gestire valori mancanti, generare espressioni CLEM, utilizzare progetti e report e assemblare stream per il deployment tramite IBM SPSS Collaboration and Deployment Services, le applicazioni predittive o IBM SPSS Modeler Advantage.
- **IBM SPSS Modeler Source, Process, and Output Nodes.** 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 Modelli.** Descrizioni di tutti i nodi utilizzati per creare modelli di data mining. IBM® SPSS® Modeler offre numerosi metodi di modellazione ricavati dall'apprendimento automatico, dall'intelligenza artificiale e dalla statistica. Per ulteriori informazioni, vedere l'argomento Panoramica sui nodi Modelli in il capitolo 3 in *IBM SPSS Modeler 15 Nodi Modelli*.
- **IBM SPSS Modeler Algorithms Guide.** Descrizione dei fondamenti di matematica per i metodi di modellazione utilizzati in SPSS Modeler. Questa guida è disponibile solo in formato PDF.

- **IBM SPSS Modeler Guida alle applicazioni.** 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, vedere l'argomento Esempi di applicazioni in *Manuale dell'utente di IBM SPSS Modeler 15*.
- **IBM SPSS Modeler Script e automazione.** Informazioni sulle modalità di automazione del sistema tramite script, incluse le proprietà che è possibile utilizzare per manipolare nodi e stream.
- **IBM SPSS Modeler Deployment Guide.** Informazioni sull'esecuzione di stream e scenari 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 (quali routine di elaborazione di dati o algoritmi di modellazione) come nodi in SPSS Modeler.
- **IBM SPSS Modeler Guida alla modellazione in-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.
- **Manuale dell'utente di IBM SPSS Modeler Administration Console.** Informazioni sull'installazione e l'utilizzo dell'interfaccia utente della console per il monitoraggio e la configurazione di SPSS Modeler Server. La console viene implementata come plug-in dell'applicazione Deployment Manager.
- **IBM SPSS Modeler Solution Publisher Guide.** SPSS Modeler Solution Publisher è un componente aggiuntivo che consente di pubblicare gli stream al di fuori dell'ambiente SPSS Modeler standard.
- **Guida CRISP-DM di IBM SPSS Modeler.** Guida passo a passo al data mining tramite la metodologia CRISP-DM con SPSS Modeler.
- **Manuale dell'utente di IBM SPSS Modeler Batch.** 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.

- **Manuale dell'utente di IBM SPSS Modeler Entity Analytics.** 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.
- **Manuale dell'utente di IBM SPSS Modeler Social Network Analysis.** Guida che spiega come eseguire l'analisi dei social network con SPSS Modeler; comprende l'analisi di gruppo e l'analisi di diffusione.

- **Manuale dell'utente di Text Analytics for SPSS Modeler.** 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.
- Manuale dell'utente di **Text Analytics for IBM SPSS Modeler Administration Console.** 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 Text Analytics for SPSS Modeler. 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 aziendale 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 gli stream di esempio sono installati nella cartella *Demos* nella directory di installazione del prodotto. Per ulteriori informazioni, vedere l'argomento Cartella Demos in *Manuale dell'utente di IBM SPSS Modeler 15*.

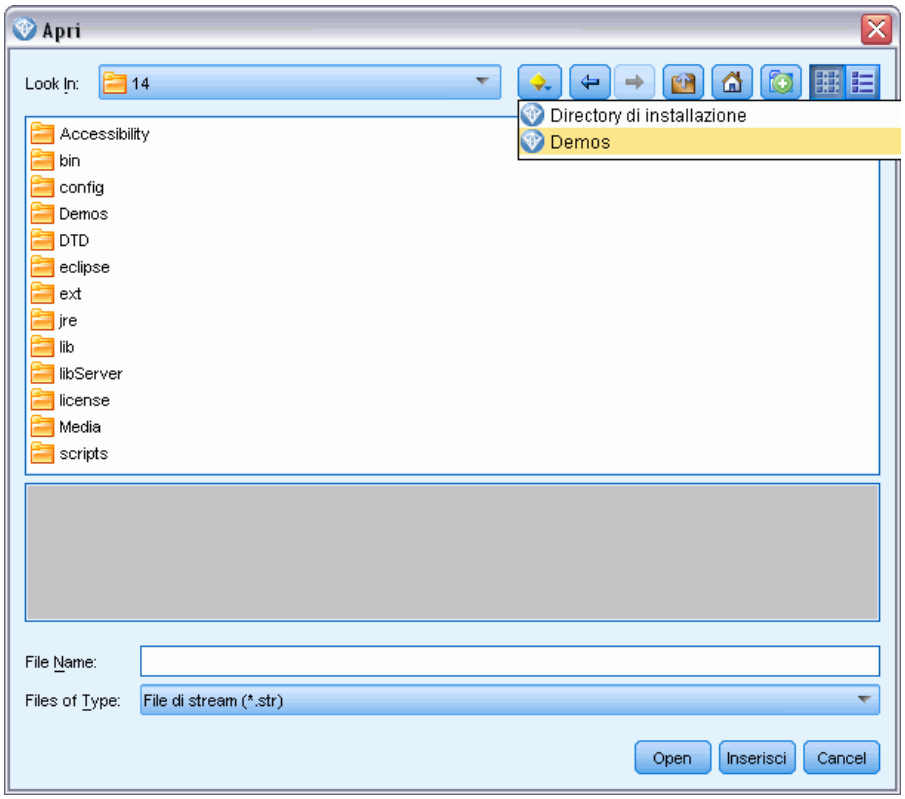
Esempi di modellazione in-database. Vedere gli esempi nella *IBM SPSS Modeler Guida alla modellazione in-database*.

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

Cartella Demos

I file di dati e gli stream di esempio utilizzati negli esempi di applicazioni sono installati nella cartella *Demos* nella directory di installazione del prodotto. A questa cartella è possibile accedere anche dal gruppo di programmi IBM SPSS Modeler 15 nel menu Start di Windows oppure facendo clic su *Demos* nell'elenco delle directory recenti nella finestra di dialogo Apri file.

Figura 1-1
Selezione della cartella Demos dall'elenco delle directory utilizzate di recente



Capitolo
2

Mining in-database

Panoramica sulla modellazione di database

IBM® SPSS® Modeler Server supporta l'integrazione con gli strumenti di data mining e di modellazione forniti da sviluppatori di database, quali IBM Netezza, IBM DB2 InfoSphere Warehouse, Oracle Data Miner 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. Ciò consente di combinare le funzionalità analitiche e la semplicità d'uso di SPSS Modeler con la potenza e le performance di un database, sfruttando gli algoritmi nativi del database distribuiti da questi fornitori. I modelli vengono creati all'interno del database e possono essere successivamente selezionati per calcolarne il punteggio attraverso l'interfaccia di SPSS Modeler secondo la procedura standard; se necessario, ne può essere eseguito il deployment attraverso IBM® SPSS® Modeler Solution Publisher. Gli algoritmi supportati sono elencati nella palette Modelli in-database di SPSS Modeler.

L'utilizzo di SPSS Modeler per accedere ad algoritmi nativi di database assicura numerosi vantaggi:

- Gli algoritmi in-database sono spesso strettamente integrati con il server di database e possono offrire performance migliorate.
- Per i modelli creati e archiviati "all'interno di database" il processo di deployment e di condivisione con tutte le applicazioni in grado di accedere al database è molto più facile da eseguire.

Generazione SQL. La modellazione in-database è distinta dalla generazione SQL, altrimenti nota come "push back SQL", che consente di generare istruzioni SQL per operazioni native di SPSS Modeler che è possibile "rinviare" al (ovvero eseguire nel) database per migliorare le prestazioni. Per esempio, i nodi Unione, Aggregazione e Seleziona generano tutti codice SQL che può essere rinviato al database per l'esecuzione. L'utilizzo della generazione SQL in combinazione con la modellazione in-database può generare stream eseguibili dall'inizio alla fine nel database, con significativi miglioramenti a livello di prestazioni rispetto agli stream eseguiti in SPSS Modeler. Per ulteriori informazioni, vedere l'argomento Ottimizzazione SQL in il capitolo 6 in *IBM SPSS Modeler Server 15 Guida della performance e amministrazione*.

Nota: le funzionalità di modellazione in-database e ottimizzazione SQL richiedono che sul computer IBM® SPSS® Modeler sia attivata la connettività SPSS Modeler Server. Con questa impostazione attivata, è possibile accedere agli algoritmi di database, restituire codice SQL direttamente da SPSS Modeler e accedere a SPSS Modeler Server. Per verificare lo stato attuale della licenza, scegliere le seguenti opzioni dal menu SPSS Modeler.

Guida > Informazioni su > Ulteriori dettagli

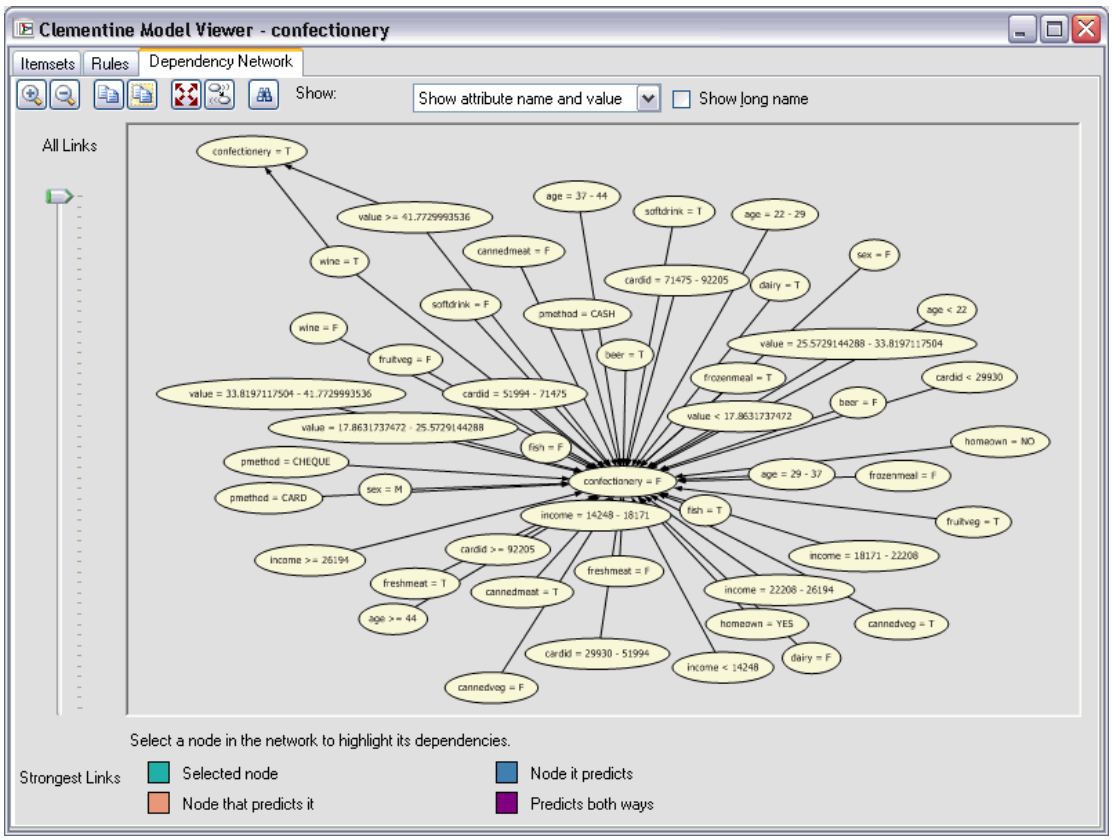
Se la connettività è abilitata, l'opzioneAbilitazione server viene visualizzata nella scheda Stato della licenza.

Per ulteriori informazioni, vedere l'argomento Connessione a IBM SPSS Modeler Server in il capitolo 3 in *Manuale dell'utente di IBM SPSS Modeler 15*.

Figura 2-1
Palette Modelli in-database

Per informazioni sugli algoritmi supportati, fare riferimento alle sezioni relative ai fornitori specifici riportate di seguito.

Figura 2-2
Visualizzatore che fornisce una visualizzazione grafica dei risultati della modellazione con regole di associazione di Microsoft Analysis Services



Requisiti necessari

Per eseguire la modellazione di database, occorre disporre di quanto elencato di seguito:

- Una connessione ODBC a un database appropriato, in cui siano installati i componenti analitici richiesti (Microsoft Analysis Services, Oracle Data Miner o IBM DB2 InfoSphere Warehouse).

- In IBM® SPSS® Modeler la modellazione di database deve essere attivata nella finestra di dialogo Applicazioni di supporto (Strumenti > Applicazioni di supporto).
- In IBM® SPSS® Modeler e in IBM® SPSS® Modeler Server (se utilizzato) le impostazioni Genera SQL e Ottimizzazione SQL devono essere attivate nella finestra di dialogo Opzioni utente. Per ulteriori informazioni, vedere l'argomento Performance/ottimizzazione in il capitolo 4 in *IBM SPSS Modeler Server 15 Guida della performance e amministrazione*. L'ottimizzazione SQL non è strettamente necessaria per la corretta operatività del processo di modellazione di database, ma è vivamente consigliata per questioni di performance.

Nota: le funzionalità di modellazione in-database e ottimizzazione SQL richiedono che sul computer SPSS Modeler sia attivata la connettività SPSS Modeler Server. Con questa impostazione attivata, è possibile accedere agli algoritmi di database, restituire codice SQL direttamente da SPSS Modeler e accedere a SPSS Modeler Server. Per verificare lo stato attuale della licenza, scegliere le seguenti opzioni dal menu SPSS Modeler.
Guida > Informazioni su > Ulteriori dettagli

Se la connettività è abilitata, l'opzioneAbilitazione server viene visualizzata nella scheda Stato della licenza.

Per ulteriori informazioni, vedere l'argomento Connessione a IBM SPSS Modeler Server in il capitolo 3 in *Manuale dell'utente di IBM SPSS Modeler 15*.

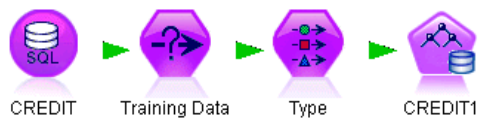
Per informazioni dettagliate, vedere le sezioni relative ai fornitori specifici riportate di seguito.

Costruzione del modello

Il processo di creazione di modelli e di calcolo del relativo punteggio mediante algoritmi di database presenta molte analogie con altri tipi di data mining all'interno di IBM® SPSS® Modeler. Il processo generale di utilizzo di nodi e "insiemi" di modelli è analogo a qualsiasi altro stream quando si lavora in SPSS Modeler. L'unica differenza è rappresentata dal fatto che l'elaborazione e la creazione di modelli effettive sono rinviate al database.

Per esempio, lo stream riportato di seguito è concettualmente identico ad altri stream di dati in SPSS Modeler. Tuttavia, esegue tutte le operazioni in un database, inclusa la creazione di modelli tramite il nodo Albero decisionale di Microsoft. Quando si esegue lo stream, SPSS Modeler fornisce al database le istruzioni necessarie per creare e archiviare il modello risultante e i dettagli vengono scaricati all'interno di SPSS Modeler.

Figura 2-3
Stream di esempio sulla modellazione di database, in cui i nodi con ombreggiatura viola indicano l'esecuzione in-database



Data Preparation

Indipendentemente dal fatto che siano utilizzati o meno algoritmi nativi di database, le preparazioni dei dati devono sempre essere rinviate al database quando possibile per migliorare le prestazioni.

- Se i dati originali sono archiviati nel database, l'obiettivo è quello di mantenerli nel database assicurandosi che tutte le operazioni a monte necessarie possano essere convertite in SQL. Questo impedisce che i dati vengano scaricati in IBM® SPSS® Modeler, evitando un collo di bottiglia che potrebbe vanificare qualsiasi vantaggio, e consentendo all'intero stream di essere eseguito nel database. Per ulteriori informazioni, vedere l'argomento Ottimizzazione SQL in il capitolo 6 in *IBM SPSS Modeler Server 15 Guida della performance e amministrazione*.
- Se i dati originali *non* sono archiviati nel database, sarà comunque possibile utilizzare la modellazione di database. In questo caso, la preparazione dei dati viene effettuata all'interno di SPSS Modeler e l'insieme dei dati preparato viene automaticamente caricato nel database per la creazione di modelli.

Calcolo del punteggio dei modelli

I modelli generati da IBM® SPSS® Modeler mediante il mining in-database differiscono dai normali modelli dell'applicazione. Sebbene vengano visualizzati nel manager Modelli come "insiemi" di modelli generati, costituiscono in effetti modelli remoti memorizzati nel server di database o di data mining remoto. Quelli visibili in SPSS Modeler sono semplicemente dei riferimenti a tali modelli remoti. In altri termini, il modello di SPSS Modeler visualizzato è un modello "vuoto", che contiene informazioni come il nome host del server di database, il nome del database e il nome del modello. Si tratta di una distinzione importante da comprendere per la visualizzazione e il calcolo del punteggio dei modelli creati utilizzando gli algoritmi nativi di database.

Figura 2-4
"Insieme" di modelli generati per alberi decisionali Microsoft

Una volta creato un nuovo modello, è possibile aggiungerlo allo stream per il calcolo del punteggio seguendo la prassi utilizzata per qualsiasi altro modello generato in SPSS Modeler. Tutti i calcoli di punteggio vengono eseguiti all'interno del database, anche se le operazioni a monte vengono eseguite altrove. (Le operazioni a monte possono essere rimandate al database per migliorare le performance, ma questo non è necessario perché avvenga il calcolo del punteggio.) Nella maggior parte dei casi, è anche possibile sfogliare il modello generato utilizzando il browser standard fornito con il database.

Per sfogliare e calcolare i punteggi, è necessario disporre di una connessione live al server su cui vengono eseguiti Oracle Data Miner, IBM DB2 InfoSphere Warehouse oppure Microsoft Analysis Services.

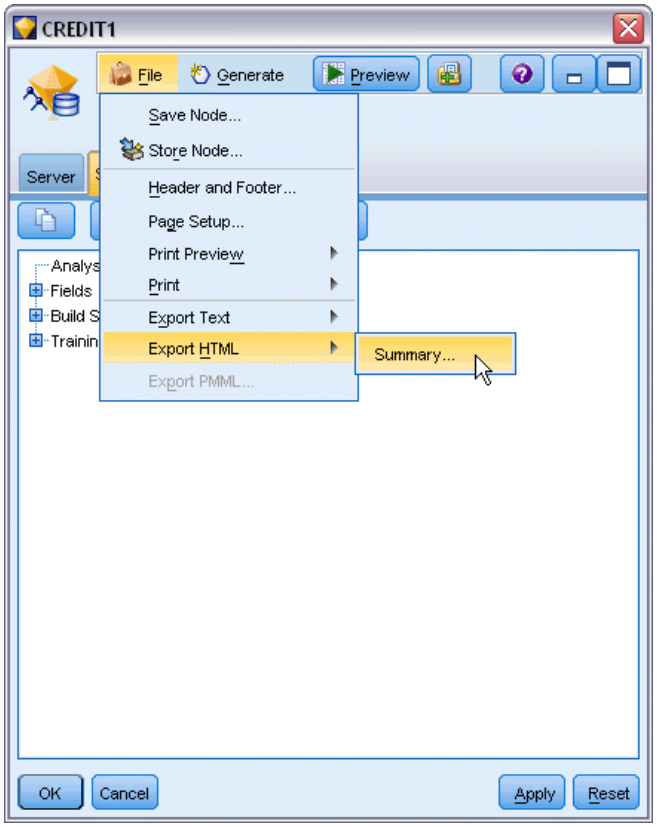
Visualizzazione dei risultati e specifica delle impostazioni

Per visualizzare i risultati e specificare le impostazioni inerenti al calcolo del punteggio, fare doppio clic sul modello nell'area di disegno dello stream. In alternativa, è possibile fare clic con il pulsante destro del mouse sul modello e scegliere Visualizza o Modifica. Le impostazioni specifiche dipendono dal tipo di modello.

Esportazione e salvataggio di modelli di database

I modelli e i riepiloghi del database possono essere esportati dal visualizzatore modelli con la stessa procedura impiegata per altri modelli creati in IBM® SPSS® Modeler, utilizzando le opzioni disponibili nel menu File.

Figura 2-5
Esportazione di un riepilogo di modello di albero decisionale Microsoft come HTML



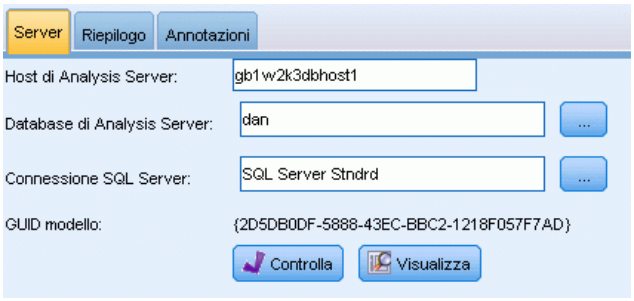
- Dal menu File del visualizzatore modelli scegliere una qualsiasi delle seguenti opzioni:
  - Esporta testo esporta il riepilogo di modello in un file di testo
  - Esporta HTML esporta il riepilogo di modello in un file HTML
  - Esporta PMML (supportata solo per i modelli IBM DB2 IM) esporta il modello come PMML (Predictive Model Markup Language), che può essere utilizzato con altri software compatibili con PMML. Per ulteriori informazioni, vedere l'argomento Importazione ed esportazione di modelli come PMML in il capitolo 10 in *Manuale dell'utente di IBM SPSS Modeler 15*.

Nota: è anche possibile salvare un modello generato scegliendo Salva nodo dal menu File. Per ulteriori informazioni, vedere l'argomento Esplorazione degli insiemi di modelli in il capitolo 3 in *IBM SPSS Modeler 15 Nodi Modelli*.

Uniformità dei modelli

Per ogni modello di database generato, IBM® SPSS® Modeler archivia una descrizione della relativa struttura insieme a un riferimento al modello con lo stesso nome memorizzato nel database. Nella scheda Server di un modello generato viene visualizzata una chiave univoca generata specificamente per il modello in questione che corrisponde al modello effettivo nel database.

Figura 2-6
Chiave del modello generato e opzioni di controllo



SPSS Modeler utilizza queste chiavi generate casualmente per controllare l'uniformità dei modelli. La chiave viene archiviata nella descrizione del modello al momento della creazione. È consigliabile verificare la corrispondenza delle chiavi prima di eseguire uno stream di deployment.

- Fare clic sul pulsante Controllo per verificare l'uniformità del modello archiviato nel database confrontando la relativa descrizione con la chiave casuale memorizzata da SPSS Modeler. Se non è possibile trovare il modello di database o la chiave non corrisponde, verrà segnalato un errore.

Visualizzazione ed esportazione di codice SQL generato

Prima di procedere all'esecuzione, è possibile visualizzare un anteprima del codice SQL, il che può essere molto utile ai fini del debug. Per ulteriori informazioni, vedere l'argomento Anteprima di SQL generato in il capitolo 6 in *IBM SPSS Modeler Server 15 Guida della performance e amministrazione*.

Capitolo
3

Modellazione di database con Microsoft Analysis Services

IBM SPSS Modeler e Microsoft Analysis Services

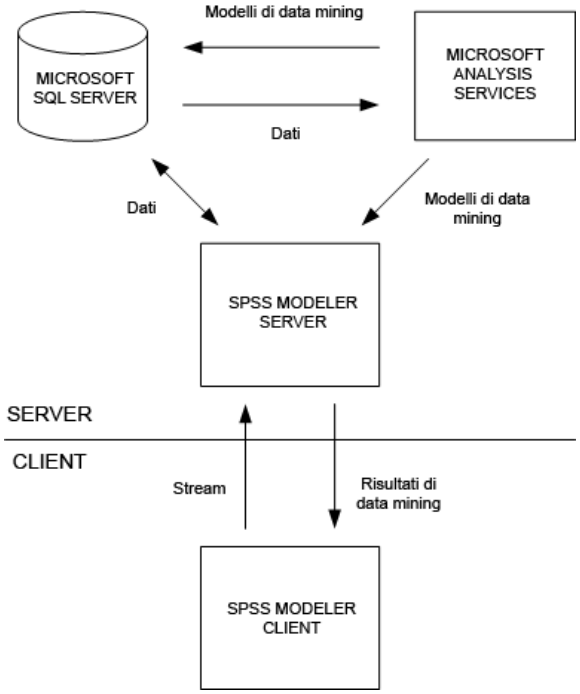
IBM® SPSS® Modeler supporta l'integrazione con Microsoft SQL Server Analysis Services. Questa funzionalità viene implementata sotto forma di nodi Modelli in SPSS Modeler ed è disponibile nella palette Modelli in-database. Se la palette non è visibile, è possibile attivarla abilitando l'integrazione con MS Analysis Services, disponibile nella scheda Microsoft della finestra di dialogo Applicazioni di supporto. Per ulteriori informazioni, vedere l'argomento Attivazione dell'integrazione con Analysis Services a pag. 17.

SPSS Modeler supporta l'integrazione con i seguenti algoritmi di Analysis Services:

- Alberi decisionali
- Raggruppamento tramite cluster
- Regole di associazione
- Bayes naive
- Regressione lineare
- Rete neurale
- Regressione logistica
- Serie storiche
- Cluster di sequenze

Nel seguente diagramma è illustrato il flusso di dati dal client verso il server nei casi in cui il mining in-database è gestito da IBM® SPSS® Modeler Server. La creazione di modelli viene eseguita mediante Analysis Services e il modello risultante è archiviato dallo stesso strumento. Un riferimento a tale modello viene conservato negli stream di SPSS Modeler. Il modello viene quindi scaricato da Analysis Services su Microsoft SQL Server o SPSS Modeler per il calcolo del punteggio.

Figura 3-1
Flusso di dati tra IBM SPSS Modeler, Microsoft SQL Server e Microsoft Analysis Services durante la creazione di modelli



Nota: non è necessario disporre di SPSS Modeler Server, sebbene sia possibile utilizzarlo. Il client IBM® SPSS® Modeler è in grado di elaborare autonomamente calcoli di mining in-database.

Requisiti per l'integrazione con Microsoft Analysis Services

Di seguito sono riportati i prerequisiti richiesti per eseguire la modellazione in-database utilizzando gli algoritmi di Analysis Services con IBM® SPSS® Modeler. Per garantire che queste condizioni vengano soddisfatte, può essere necessario consultare l'amministratore di sistema.

- Esecuzione di IBM® SPSS® Modeler in un'installazione di IBM® SPSS® Modeler Server (modalità distribuita) su Windows. Le piattaforme UNIX non sono supportate in questa integrazione con Analysis Services.
  Importante: gli utenti di SPSS Modeler devono configurare una connessione ODBC utilizzando il driver SQL Native Client disponibile sul sito Web di Microsoft all'indirizzo riportato di seguito in *Requisiti aggiuntivi di SPSS Modeler Server*. *Il driver fornito con IBM® SPSS® Data Access Pack, sebbene generalmente consigliato per altri usi con SPSS Modeler, non è raccomandato per questo scopo.* È necessario configurare il driver per l'uso di SQL Server con l'opzione Autenticazione integrata di Windows attivata, poiché SPSS Modeler non supporta l'autenticazione SQL Server. Per domande sulla creazione o l'impostazione di autorizzazioni per sorgenti dati ODBC, rivolgersi all'amministratore del database.
- È necessario aver installato sul computer SQL Server 2005 o 2008, sebbene non necessariamente sullo stesso host di SPSS Modeler. Gli utenti di SPSS Modeler devono disporre delle autorizzazioni richieste per leggere e scrivere dati nonché per creare ed eliminare tabelle e visualizzazioni.

Nota: si consiglia l'uso di SQL Server Enterprise Edition. La versione Enterprise Edition offre una flessibilità maggiore fornendo parametri avanzati che consentono di perfezionare i risultati degli algoritmi. La versione Standard Edition fornisce gli stessi parametri ma non consente agli utenti di modificare alcuni dei parametri avanzati.

- È necessario aver installato Microsoft SQL Server Analysis Services sullo stesso host di SQL Server.

Requisiti aggiuntivi di IBM SPSS Modeler Server

Per utilizzare gli algoritmi di Analysis Services con SPSS Modeler Server, è necessario aver installato sull'host di SPSS Modeler Server i seguenti componenti

Nota: se SQL Server è installato sullo stesso host di SPSS Modeler Server, tali componenti saranno già disponibili.

- Microsoft .NET Framework Redistributable Package versione 2.0 (x86)
- Microsoft Core XML Services (MSXML) 6.0
- Provider OLE DB Microsoft SQL Server 2008 Analysis Services 10.0 (avere cura di selezionare la versione corretta per il proprio sistema operativo)
- Microsoft SQL Server 2008 Native Client (avere cura di selezionare la versione corretta per il proprio sistema operativo)

Per scaricare questi componenti, accedere a *www.microsoft.com/downloads*, cercare .NET Framework o (per tutti gli altri componenti) SQL Server Feature Pack e selezionare il pacchetto più recente per la propria versione di SQL Server.

L'esecuzione di tali componenti potrebbe richiedere l'installazione di altri pacchetti, che dovrebbero essere disponibili anch'essi nell'area Download del sito Web di Microsoft.

Requisiti aggiuntivi di IBM SPSS Modeler

Per utilizzare gli algoritmi di Analysis Services con SPSS Modeler, è necessario che siano installati gli stessi componenti riportati in precedenza, con l'aggiunta dei seguenti sul client:

- Microsoft SQL Server 2008 Datamining Viewer Controls (avere cura di selezionare la versione corretta per il proprio sistema operativo), che richiede inoltre:
- Microsoft ADOMD.NET

Per scaricare questi componenti, accedere a *www.microsoft.com/downloads*, cercare SQL Server Feature Pack e selezionare il pacchetto più recente per la propria versione di SQL Server.

Nota: le funzionalità di modellazione in-database e ottimizzazione SQL richiedono che sul computer SPSS Modeler sia attivata la connettività SPSS Modeler Server. Con questa impostazione attivata, è possibile accedere agli algoritmi di database, restituire codice SQL direttamente da SPSS Modeler e accedere a SPSS Modeler Server. Per verificare lo stato attuale della licenza, scegliere le seguenti opzioni dal menu SPSS Modeler.

Guida > Informazioni su > Ulteriori dettagli

Se la connettività è abilitata, l'opzioneAbilitazione server viene visualizzata nella scheda Stato della licenza.

Per ulteriori informazioni, vedere l'argomento Connessione a IBM SPSS Modeler Server in il capitolo 3 in *Manuale dell'utente di IBM SPSS Modeler 15*.

Attivazione dell'integrazione con Analysis Services

Per attivare l'integrazione di IBM® SPSS® Modeler con Analysis Services, è necessario configurare SQL Server e Analysis Services e creare una sorgente ODBC, quindi attivare l'integrazione nella finestra di dialogo di Applicazioni di supporto SPSS Modeler e, infine, attivare la generazione e l'ottimizzazione SQL.

Nota: è necessario disporre di Microsoft SQL Server e di Microsoft Analysis Services. Per ulteriori informazioni, vedere l'argomento Requisiti per l'integrazione con Microsoft Analysis Services a pag. 15.

Configurazione di SQL Server

Configurare SQL Server in modo da consentire che il calcolo del punteggio sia eseguito all'interno del database.

- Creare la seguente chiave del Registro di sistema sul computer host SQL Server:

  HKEY_LOCAL_MACHINE\SOFTWARE\Microsoft\MSSQLServer\Providers\MSOLAP

- Aggiungere quindi alla chiave il seguente valore DWORD:

  AllowInProcess 1

- Riavviare SQL Server dopo aver apportato la modifica.

Configurazione di Analysis Services

Prima che SPSS Modeler possa comunicare con Analysis Services, è necessario configurare manualmente due impostazioni nella finestra di dialogo Proprietà di Analysis Server:

- Accedere ad Analysis Server tramite MS SQL Server Management Studio.
- Accedere alla finestra di dialogo Proprietà facendo clic con il pulsante destro sul nome del server e scegliendo Proprietà.
- Selezionare la casella di controllo Mostra proprietà (tutte) avanzate.
- Modificare le seguenti proprietà:
  - Modificare il valore di `DataMining\AllowAdHocOpenRowsetQueries` su `True` (il valore di default è `False`).
  - Modificare il valore di `DataMining\AllowProvidersInOpenRowset` con `[all]` (non esiste un valore di default).

Creazione di un DSN ODBC per SQL Server

Per leggere o scrivere su un database, occorre che un'origine dati ODBC sia installata e configurata per il database in questione, con le relative autorizzazioni di lettura e scrittura. È necessario disporre del driver ODBC Microsoft SQL Native Client che viene installato automaticamente con SQL Server. *Il driver fornito con IBM® SPSS® Data Access Pack, sebbene generalmente consigliato per altri usi con SPSS Modeler, non è raccomandato per questo scopo.* Se SPSS Modeler e SQL Server risiedono su host diversi, è possibile scaricare il driver ODBC Microsoft SQL Native Client. Per ulteriori informazioni, vedere l'argomento Requisiti per l'integrazione con Microsoft Analysis Services a pag. 15.

Per domande sulla creazione o l'impostazione di autorizzazioni per sorgenti dati ODBC, rivolgersi all'amministratore del database.

- Con il driver ODBC Microsoft SQL Native Client, creare un DSN ODBC che punta al database SQL Server utilizzato nel processo di data mining. Per le restanti impostazioni del driver, è necessario utilizzare le impostazioni di default.
- Assicurarsi che per il DSN sia selezionata l'autenticazione integrata di Windows.
  - Se IBM® SPSS® Modeler e IBM® SPSS® Modeler Server sono in esecuzione su host diversi, creare lo stesso DSN ODBC su ogni host. Assicurarsi di utilizzare lo stesso nome DSN su ogni host.

Attivazione dell'integrazione di Analysis Services in IBM SPSS Modeler

Per consentire a SPSS Modeler di utilizzare Analysis Services, è innanzitutto necessario specificare le informazioni sul server nella finestra di dialogo Applicazioni di supporto.

- Dai menu di SPSS Modeler scegliere:
  Strumenti > Opzioni > Applicazioni di supporto
- Fare clic sulla scheda Microsoft.
  - **Attiva integrazione di Microsoft Analysis Services.** Attiva la palette Modelli in-database (se non è già visualizzata) nella parte inferiore della finestra SPSS Modeler e aggiunge i nodi degli algoritmi di Analysis Services.

Figura 3-2
Scheda Modelli database

  - **Host di Analysis Server.** Specificare il nome del computer su cui è in esecuzione Analysis Services.
  - **Database di Analysis Server.** Selezionare il database desiderato facendo clic sul pulsante con i puntini di sospensione (...) che consente di aprire una sottofinestra di dialogo in cui è possibile scegliere tra i database disponibili. L'elenco contiene i database disponibili per il server Analysis specificato. Poiché Microsoft Analysis Services archivia i modelli di data

mining all'interno di database denominati, è necessario selezionare il database appropriato in cui vengono archiviati i modelli Microsoft creati da SPSS Modeler.

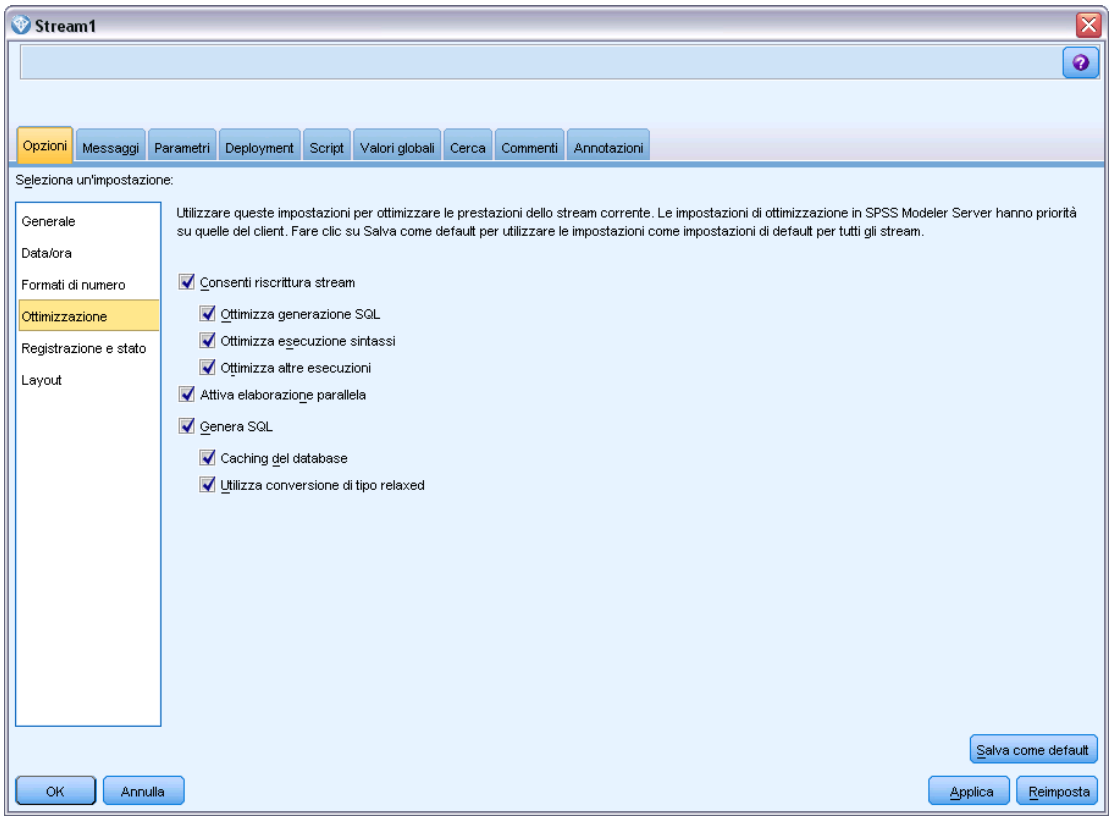
- **Connessione SQL Server.** Specificare le informazioni DSN utilizzate dal database SQL Server per archiviare i dati passati ad Analysis Server. Scegliere la sorgente dati ODBC che verrà utilizzata per fornire i dati necessari per la creazione di modelli di data mining Analysis Services. Se si creano modelli Analysis Services a partire da dati forniti all'interno di file piatti o sorgenti dati ODBC, i dati verranno automaticamente caricati in una tabella temporanea creata nel database SQL Server al quale punta la sorgente dati ODBC.
- **Avvisa prima di sovrascrivere un modello di data mining.** Selezionare questa opzione per assicurarsi che i modelli archiviati nel database non vengano sovrascritti da SPSS Modeler senza preavviso.

Nota: Le impostazioni specificate nella finestra di dialogo Applicazioni di supporto possono essere sovrascritte all'interno di vari nodi di Analysis Services.

Attivazione di generazione e ottimizzazione SQL

▶ Dai menu di SPSS Modeler scegliere:
Strumenti > Proprietà stream > Opzioni

Figura 3-3
Impostazioni di ottimizzazione



▶ Fare clic sull'opzione Ottimizzazione nel riquadro di spostamento.

- Confermare che l'opzione Genera SQL è attivata. Questa impostazione è necessaria per il corretto funzionamento della modellazione di database.
- Selezionare Ottimizza generazione SQL e Ottimizza altre esecuzioni (queste due opzioni non sono strettamente necessarie, tuttavia se ne consiglia la selezione per ottenere performance ottimizzate).

Per ulteriori informazioni, vedere l'argomento Impostazione delle opzioni di ottimizzazione per gli stream in il capitolo 5 in *Manuale dell'utente di IBM SPSS Modeler 15*.

Creazioni di modelli con Analysis Services

La creazione di modelli di Analysis Services richiede che l'insieme di dati addestramento sia posizionato in una tabella o visualizzazione all'interno del database SQL Server. Se i dati non sono ubicati in SQL Server o devono essere elaborati in IBM® SPSS® Modeler come parte del processo di preparazione dei dati che non è possibile eseguire in SQL Server, tali dati vengono automaticamente caricati in una tabella temporanea di SQL Server prima della creazione dei modelli.

Gestione di modelli di Analysis Services

La creazione di un modello di Analysis Services tramite IBM® SPSS® Modeler comporta la creazione di un modello in SPSS Modeler e la creazione o la sostituzione di un modello nel database SQL Server. Il modello di SPSS Modeler fa riferimento al contenuto di un modello di database archiviato in un server di database. SPSS Modeler consente di eseguire un controllo dell'uniformità archiviando una stringa identica con la chiave del modello generato sia nel modello SQL Server che nel modello di SPSS Modeler.

Il nodo Modelli **Albero decisionale MS** è utilizzato nella modellazione predittiva di attributi sia categoriali che continui. Per gli attributi categoriali, il nodo esegue previsioni in base alle relazioni tra le colonne di input in un insieme di dati. Per esempio, in uno scenario per prevedere quali clienti è probabile che acquistino una bicicletta, se nove su dieci clienti più giovani acquistano una bicicletta, ma solo due su dieci clienti più anziani la acquistano, il nodo desume che l'età sia un buon predittore dell'acquisto di biciclette. L'albero decisionale esegue previsioni in base a questa tendenza verso un particolare risultato. Per gli attributi continui, l'algoritmo utilizza la regressione lineare per stabilire dove l'albero decisione si suddivide. Se più di una colonna è impostata come prevedibile, o se i dati di input contengono una tabella nidificata che è impostata come prevedibile, il nodo genera un albero decisionale separato per ogni colonna prevedibile.

Il nodo Modelli **Raggruppamento cluster MS** utilizza tecniche iterative per raggruppare i casi di un insieme di dati in cluster contenenti caratteristiche simili. Questi raggruppamenti sono utili per l'esplorazione dei dati, l'individuazione di anomalie nei dati e la creazione di previsioni. I modelli di raggruppamento tramite cluster individuano le relazioni di un insieme di dati che non potrebbero essere derivate logicamente dall'osservazione casuale. Per esempio, è possibile comprendere logicamente che le persone che si recano al lavoro in bicicletta in genere non abitano molto distante dal posto di lavoro. Tuttavia, l'algoritmo è in grado di trovare altre caratteristiche relative ai pendolari della bicicletta che non sono così ovvie. Il nodo di raggruppamento cluster differisce dagli altri nodi di data mining in quanto non è specificato alcun campo obiettivo. Il nodo di raggruppamento cluster addestra il modello partendo strettamente dalla relazione esistente nei dati e dai cluster identificati dal nodo.

Il nodo Modelli **Regole di associazione MS** è utile per i motori di raccomandazioni. Un motore di raccomandazioni consiglia i prodotti ai clienti in base agli elementi già acquistati o per i quali hanno mostrato un interesse. I modelli di associazione vengono costruiti sulla base di insiemi di dati che contengono identificatori sia per i singoli casi che per gli elementi contenuti nei casi. Un gruppo di elementi di un caso viene definito **insieme di elementi**. Un modello di associazione è costituito da una serie di insiemi di elementi e dalle regole che descrivono come questi elementi sono raggruppati all'interno dei casi. Le regole individuate dall'algoritmo possono essere utilizzate per prevedere i probabili acquisti futuri di un cliente, in base agli elementi già presenti nel suo carrello.

Il nodo Modelli **Bayes naive MS** calcola la probabilità condizionale tra i campi obiettivo e predittore e presume che le colonne siano indipendenti. Il modello viene definito naïve perché considera tutte le variabili di previsione proposte come indipendenti l'una dall'altra. Questo metodo è meno intenso dal punto di vista computazionale rispetto agli altri algoritmi Analysis Services e pertanto è utile per scoprire rapidamente le relazioni durante le fasi preliminari di modellazione. Questo nodo può essere utile per effettuare esplorazioni iniziali dei dati e successivamente applicare i risultati per creare modelli aggiuntivi con altri nodi che possono richiedere un tempo di calcolo più lungo ma fornire risultati più precisi.

Il nodo Modelli **Regressione lineare MS** è una variazione del nodo Alberi decisionali, dove il parametro MINIMUM_LEAF_CASES è impostato in modo da essere maggiore o uguale al numero totale dei casi nell'insieme di dati che il nodo utilizza per addestrare il modello di mining. Con il suddetto parametro impostato in questo modo, il nodo non creerà mai una suddivisione e verrà pertanto eseguita una regressione lineare.

Il nodo Modelli **Rete neurale MS** è simile al nodo Albero decisionale MS, poiché calcola le probabilità per ogni possibile stato dell'attributo di input quando viene fornito ogni stato dell'attributo prevedibile. È quindi possibile utilizzare queste probabilità in un momento successivo per prevedere un risultato dell'attributo previsto, in base agli attributi di input.

Il nodo Modelli **Regressione logistica MS** è una variazione del nodo Rete neurale MS, dove il parametro HIDDEN_NODE_RATIO è impostato su 0. Questo parametro crea un modello di rete neurale che non contiene uno strato nascosto ed è pertanto equivalente alla regressione logistica.

Il nodo Modelli **Serie storica MS** prevede degli algoritmi di regressione ottimizzati per la previsione di valori continui nel tempo, per esempio le vendite di un prodotto. A differenza di altri algoritmi Microsoft (quali gli alberi decisionali), un modello di serie storica non richiede colonne aggiuntive di nuove informazioni come input per prevedere una tendenza. I modelli di serie storica possono infatti prevedere le tendenze solo in base all'insieme di dati originale utilizzato per creare il modello. È possibile anche aggiungere nuovi dati al modello quando si effettua una previsione e incorporare automaticamente i nuovi dati nell'analisi della tendenza. Per ulteriori informazioni, vedere l'argomento Nodo Serie storica MS a pag. 32.

Il nodo Modelli **Cluster di sequenze MS** identifica le sequenze ordinate presenti nei dati e combina i risultati di questa analisi con le tecniche di raggruppamento tramite cluster per generare cluster basati sulle sequenze e su altri attributi. Per ulteriori informazioni, vedere l'argomento Nodo Cluster di sequenze MS a pag. 35.

È possibile accedere a ogni nodo dalla palette Modelli database nella parte inferiore della finestra di SPSS Modeler.

Impostazioni comuni a tutti i nodi degli algoritmi

Le seguenti impostazioni sono valide per tutti gli algoritmi di Analysis Services.

Opzioni server

Nella scheda Server è possibile configurare l'host e il database di Analysis Server nonché la sorgente dati SQL Server. Le opzioni specificate in questa posizione sovrascrivono quelle selezionate nella scheda Microsoft all'interno della finestra di dialogo Applicazioni di supporto. Per ulteriori informazioni, vedere l'argomento Attivazione dell'integrazione con Analysis Services a pag. 17.

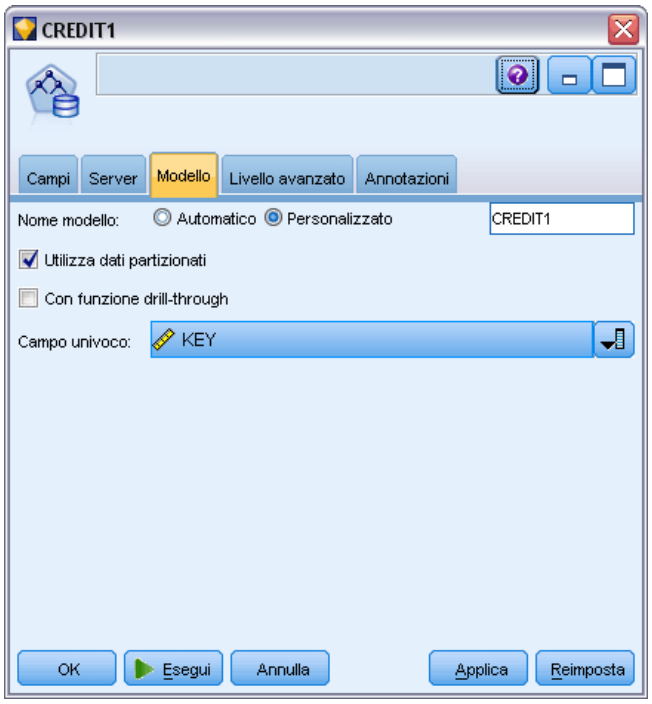
Figura 3-4
Opzioni server

Nota: Una variante di questa scheda è inoltre disponibile quando si calcola il punteggio di modelli di Analysis Services. Per ulteriori informazioni, vedere l'argomento Scheda Server dell'insieme di modelli Analysis Services a pag. 39.

Opzioni modello

Per poter creare il modello di base, occorre specificare preliminarmente una serie di opzioni nella scheda Modello. Il metodo per il calcolo del punteggio e altre opzioni avanzate sono accessibili nella scheda Livello avanzato.

Figura 3-5
Opzioni modello



Di seguito sono riportate le principali opzioni di modellazione disponibili:

Nome modello. Specifica il nome assegnato al modello creato 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 modello creato.

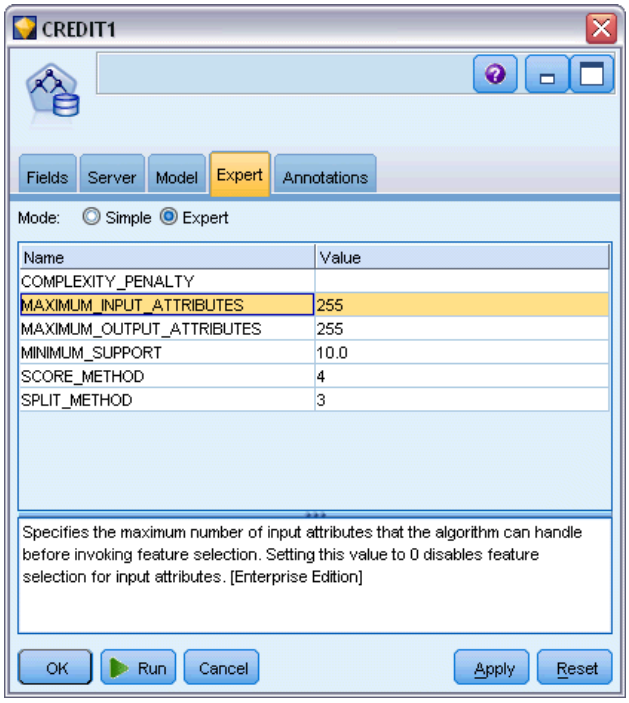
Utilizza dati partizionati. Suddivide i dati in sottoinsiemi separati, o campioni, per le fasi di addestramento, test e convalida in base al campo di partizione corrente. Utilizzando un campione per creare il modello e un altro campione per sottoporlo a verifica, è possibile ottenere una valida indicazione del modo in cui il modello potrà essere esteso a insiemi di dati di dimensioni maggiori, più simili ai dati correnti. Se non viene specificato un campo di partizione nello stream, tale opzione viene ignorata. Per ulteriori informazioni, vedere l'argomento Nodo Partizione in il capitolo 4 in *IBM SPSS Modeler 15 Nodi di input, elaborazione e output*.

Con funzione drill-through. Se visualizzata, questa opzione consente di interrogare il modello per ottenere informazioni sui casi compresi nel modello.

Campo univoco. Dall'elenco a discesa,selezionare un campo che identifichi in modo univoco ogni caso. In genere, si tratta di un campo ID, per esempio IDCliente.

Opzioni avanzate Albero decisionale MS

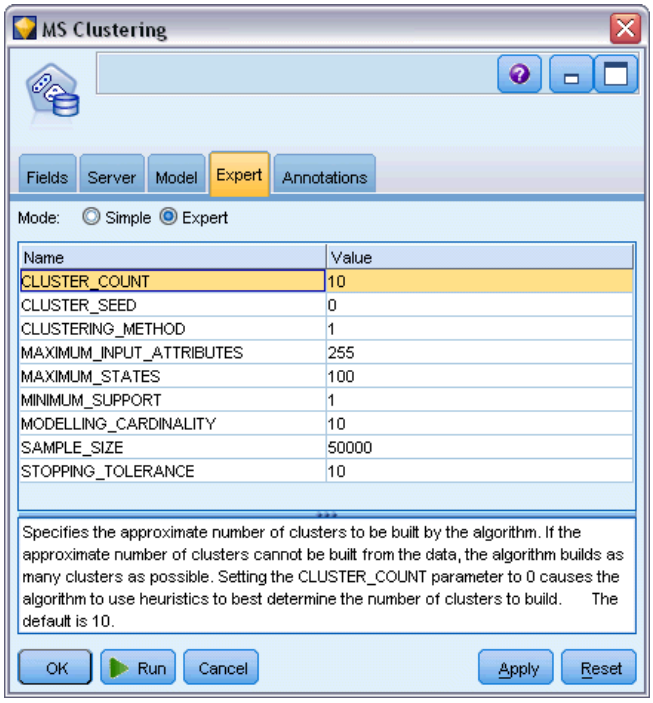
Figura 3-6
Opzioni avanzate Albero decisionale MS



Le opzioni disponibili nella scheda Livello avanzato possono variare a seconda della struttura dello stream selezionato. Per ulteriori dettagli sulle opzioni di livello avanzato disponibili per il nodo del modello Analysis Services selezionato, fare riferimento alla guida a livello di campo dell'interfaccia utente.

Opzioni avanzate Raggruppamento cluster MS

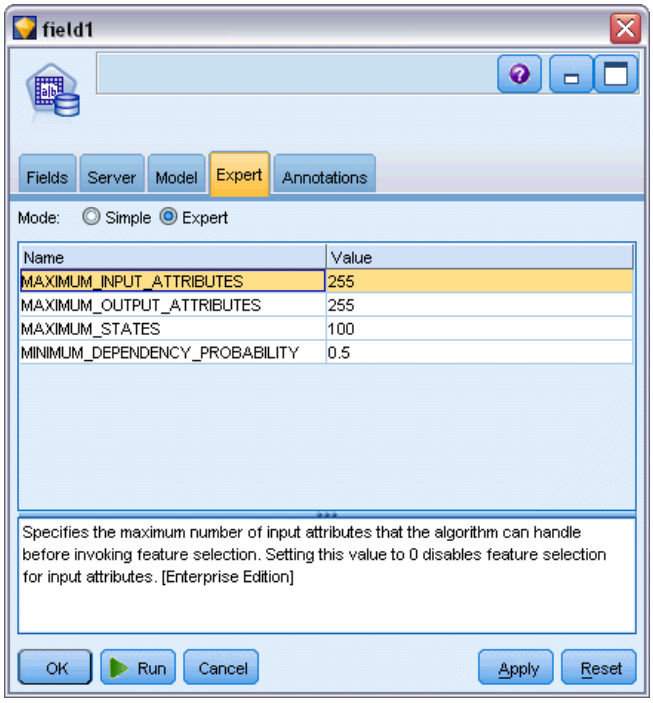
Figura 3-7
Opzioni avanzate Raggruppamento cluster MS



Le opzioni disponibili nella scheda Livello avanzato possono variare a seconda della struttura dello stream selezionato. Per ulteriori dettagli sulle opzioni di livello avanzato disponibili per il nodo del modello Analysis Services selezionato, fare riferimento alla guida a livello di campo dell'interfaccia utente.

Opzioni avanzate Bayes naive MS

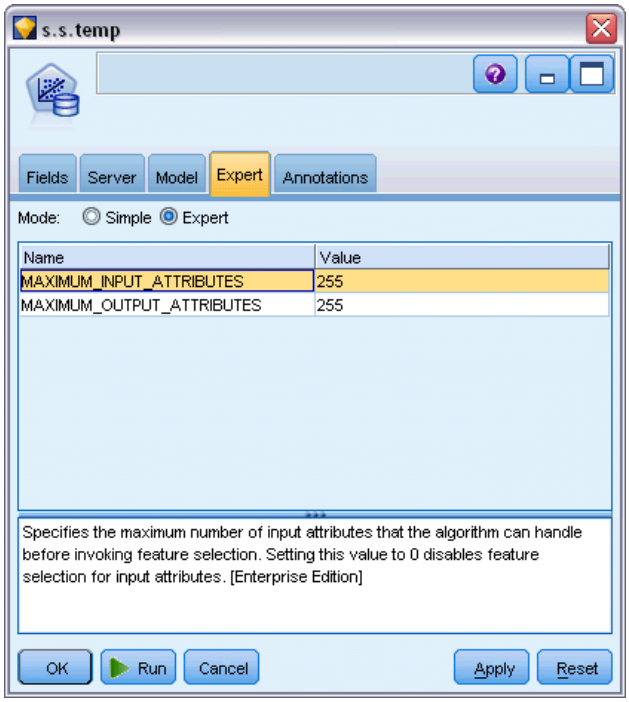
Figura 3-8
Opzioni avanzate Bayes naive MS



Le opzioni disponibili nella scheda Livello avanzato possono variare a seconda della struttura dello stream selezionato. Per ulteriori dettagli sulle opzioni di livello avanzato disponibili per il nodo del modello Analysis Services selezionato, fare riferimento alla guida a livello di campo dell'interfaccia utente.

Opzioni avanzate Regressione lineare MS

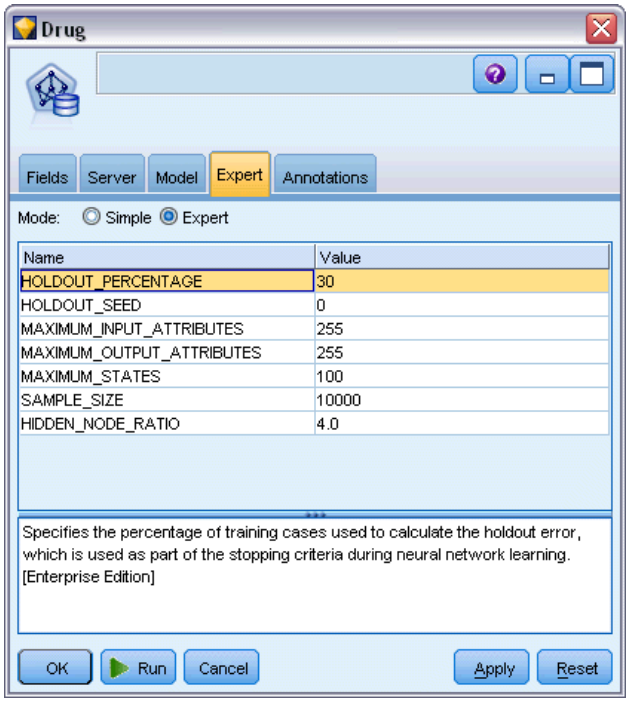
Figura 3-9
Opzioni avanzate Regressione lineare MS



Le opzioni disponibili nella scheda Livello avanzato possono variare a seconda della struttura dello stream selezionato. Per ulteriori dettagli sulle opzioni di livello avanzato disponibili per il nodo del modello Analysis Services selezionato, fare riferimento alla guida a livello di campo dell'interfaccia utente.

Opzioni avanzate Rete neurale MS

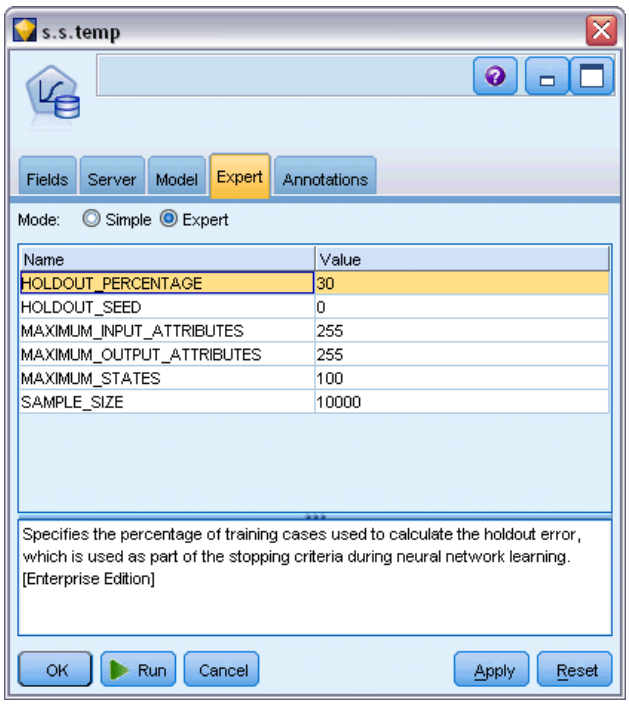
Figura 3-10
Opzioni avanzate Rete neurale MS



Le opzioni disponibili nella scheda Livello avanzato possono variare a seconda della struttura dello stream selezionato. Per ulteriori dettagli sulle opzioni di livello avanzato disponibili per il nodo del modello Analysis Services selezionato, fare riferimento alla guida a livello di campo dell'interfaccia utente.

Opzioni avanzate Regressione logistica MS

Figura 3-11
Opzioni avanzate Regressione logistica MS



Le opzioni disponibili nella scheda Livello avanzato possono variare a seconda della struttura dello stream selezionato. Per ulteriori dettagli sulle opzioni di livello avanzato disponibili per il nodo del modello Analysis Services selezionato, fare riferimento alla guida a livello di campo dell'interfaccia utente.

Nodo Regole di associazione MS

Il nodo Modelli Regole di associazione Microsoft è utile per i motori di raccomandazioni. Un motore di raccomandazioni consiglia i prodotti ai clienti in base agli elementi già acquistati o per i quali hanno mostrato un interesse. I modelli di associazione vengono costruiti sulla base di insiemi di dati che contengono identificatori sia per i singoli casi che per gli elementi contenuti nei casi. Un gruppo di elementi di un caso viene definito **insieme di elementi**.

Un modello di associazione è costituito da una serie di insiemi di elementi e dalle regole che descrivono come questi elementi sono raggruppati all'interno dei casi. Le regole individuate dall'algoritmo possono essere utilizzate per prevedere i probabili acquisti futuri di un cliente, in base agli elementi già presenti nel suo carrello.

Per i dati in formato tabulare, l'algoritmo crea punteggi che rappresentano la probabilità ($MP-*campo*) per ogni raccomandazione generata ($M-*campo*). Per i dati in formato transazionale vengono creati punteggi per supporto ($MS-*campo*), probabilità ($MP-*campo*) e probabilità regolata ($MAP-*campo*) per ogni raccomandazione generata ($M-*campo*). Per ulteriori

informazioni, vedere l'argomento Dati tabulari e dati transazionali in il capitolo 12 in *IBM SPSS Modeler 15 Nodi Modelli*.

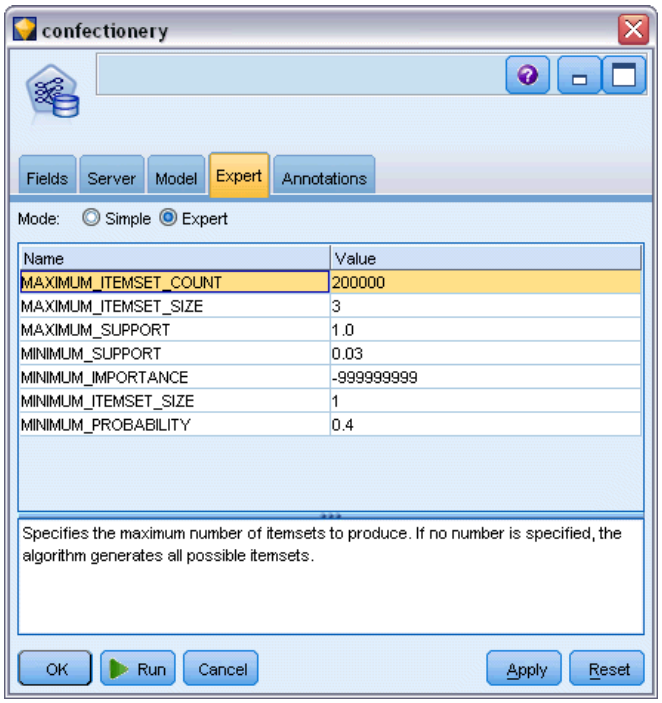
Requisiti

I requisiti di un modello di associazione transazionale sono i seguenti:

- **Campo univoco.** Un modello di regole di associazione richiede una chiave che identifichi i record in modo univoco.
- **Campo ID.** Quando si crea un modello Regole di associazione MS con dati in formato transazionale è necessario un campo ID che identifichi le singole transazioni. I campi ID si possono impostare sullo stesso valore del campo univoco.
- **Almeno un campo di input.** L'algoritmo della regola di associazione richiede almeno un campo di input.
- **Campo obiettivo.** Quando si crea un modello Regole di associazione MS con dati transazionali, il campo obiettivo deve essere il campo della transazione, per esempio i prodotti acquistati da un utente.

Opzioni avanzate Regole di associazione MS

Figura 3-12
Opzioni avanzate Regole di associazione MS



Le opzioni disponibili nella scheda Livello avanzato possono variare a seconda della struttura dello stream selezionato. Per ulteriori dettagli sulle opzioni di livello avanzato disponibili per il nodo del modello Analysis Services selezionato, fare riferimento alla guida a livello di campo dell'interfaccia utente.

Nodo Serie storica MS

Il nodo Modelli Serie storica MS supporta due tipi di previsioni:

- futura
- storica

Le **previsioni future** stimano i valori del campo obiettivo per un numero specificato di periodi di tempo successivi alla fine dei dati storici e vengono sempre eseguite. Le **previsioni storiche** sono valori stimati del campo obiettivo relativi a un numero specifico di periodi di tempo i cui valori sono effettivamente presenti nei dati storici. Le previsioni storiche si possono utilizzare per valutare la qualità del modello confrontando i dati storici effettivi con quelli previsti. Il valore del punto di partenza delle previsioni determina l'esecuzione o meno delle previsioni storiche.

A differenza del nodo Serie storica IBM® SPSS® Modeler, il nodo Serie storica MS non deve essere preceduto da un nodo Intervalli di tempo. Un'ulteriore differenza è il fatto che per default i punteggi sono calcolati solo per le righe previste e non per tutte le righe di dati storici della serie storica.

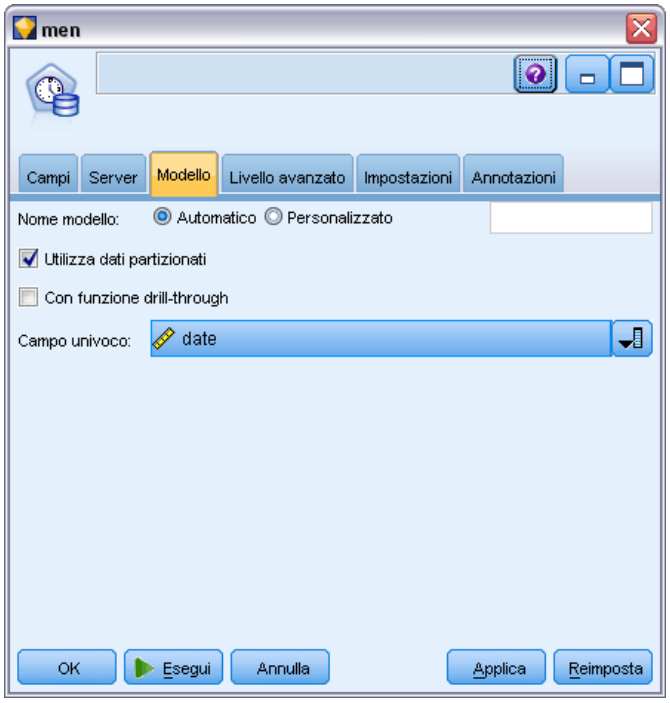
Requisiti

I requisiti di un modello Serie storica MS sono i seguenti:

- **Un solo campo chiave ora.** Ogni modello deve contenere un campo numerico o data utilizzato come serie del caso, che definisce le fasce temporali utilizzate dal modello. Il tipo di dati del campo chiave ora può essere data/ora o numerico. Tuttavia, il campo deve contenere valori continui che devono inoltre essere univoci per ogni serie.
- **Un solo campo obiettivo.** È possibile specificare un solo campo obiettivo in ogni modello. Il tipo di dati del campo obiettivo deve avere valori continui. Per esempio, è possibile prevedere come variano nel tempo attributi numerici quali il reddito, le vendite o la temperatura. Non è invece consentito l'utilizzo di un campo contenente valori categoriali quali lo stato di acquisto o il livello di istruzione come campo obiettivo.
- **Almeno un campo di input.** L'algoritmo Serie storica MS richiede almeno un campo di input. Il tipo di dati del campo di input deve avere valori continui. I campi di input non continui vengono ignorati al momento della creazione del modello.
- **L'insieme di dati deve essere ordinato.** L'insieme di dati di input deve essere ordinato (sul campo chiave ora), altrimenti la creazione del modello si interromperà con un errore.

Opzioni del modello di serie storica MS

Figura 3-13
Opzioni del modello di serie storica MS



Nome modello. Specifica il nome assegnato al modello creato 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 modello creato.

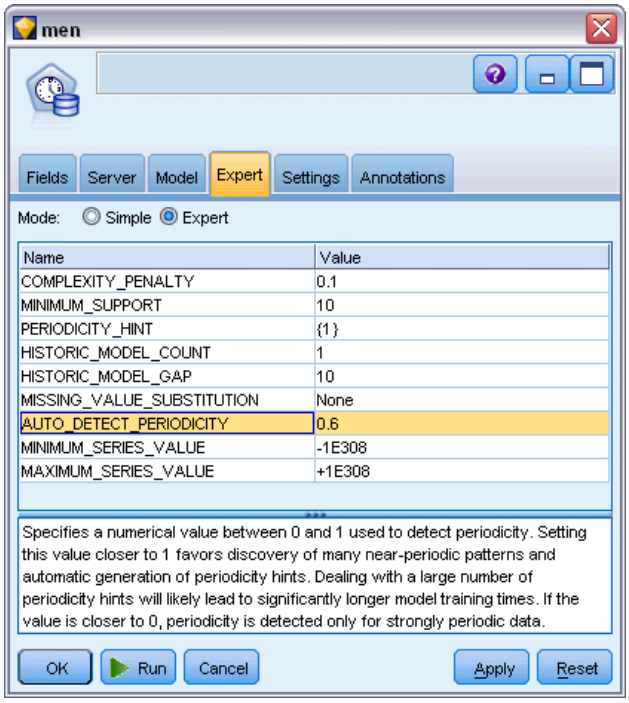
Utilizza dati partizionati. Se è definito un campo di partizione, questa opzione garantisce che per la creazione del modello verranno utilizzati solo i dati della partizione di addestramento. Per ulteriori informazioni, vedere l'argomento Nodo Partizione in il capitolo 4 in *IBM SPSS Modeler 15 Nodi di input, elaborazione e output*.

Con funzione drill-through. Se visualizzata, questa opzione consente di interrogare il modello per ottenere informazioni sui casi compresi nel modello.

Campo univoco. Dall'elenco a discesa, selezionare il campo chiave ora utilizzato per creare il modello di serie storica.

Opzioni avanzate Serie storica MS

Figura 3-14
Opzioni avanzate Serie storica MS

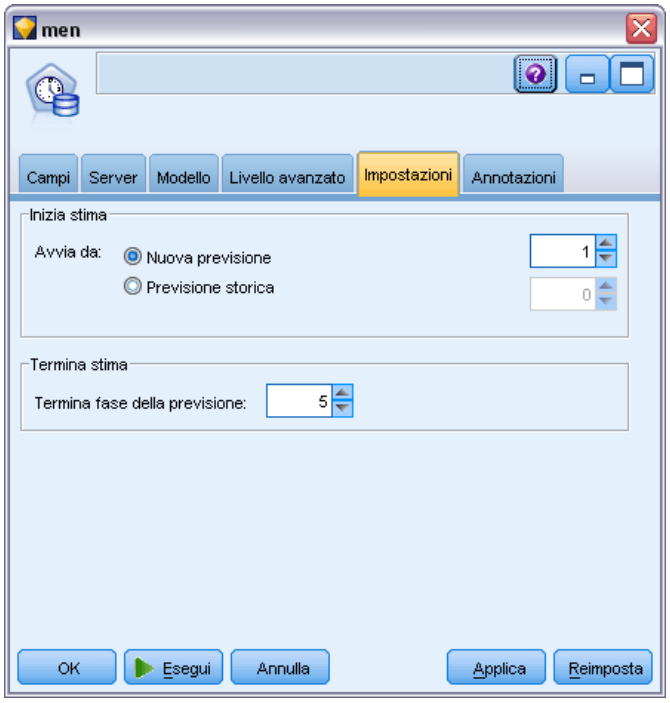


Le opzioni disponibili nella scheda Livello avanzato possono variare a seconda della struttura dello stream selezionato. Per ulteriori dettagli sulle opzioni di livello avanzato disponibili per il nodo del modello Analysis Services selezionato, fare riferimento alla guida a livello di campo dell'interfaccia utente.

In caso di previsioni storiche, il numero di fasi storiche che è possibile includere nel risultato del calcolo del punteggio viene deciso dal valore di (HISTORIC_MODEL_COUNT * HISTORIC_MODEL_GAP). Per default, questo limite è 10, ovvero vengono effettuate solo 10 previsioni storiche. In questo caso, per esempio, si verifica un errore se si inserisce una valore inferiore a -10 per Previsione storica nella scheda Impostazioni dell'insieme di modelli (vedere Scheda Impostazioni dell'insieme di modelli Serie storica MS a pag. 45). Per visualizzare un numero maggiore di previsioni storiche è possibile aumentare il valore di HISTORIC_MODEL_COUNT o HISTORIC_MODEL_GAP, ma questo aumenterà il tempo impiegato per la creazione del modello.

Opzioni di impostazione Serie storica MS

Figura 3-15
Opzioni di impostazione Serie storica MS



Inizia stima. Specificare il periodo da cui si desidera iniziare le previsioni.

- **Avvia da: Nuova previsione.** Il periodo da cui si desidera far iniziare le previsioni future, espresso come distanza dall'ultimo periodo dei dati storici di cui si dispone. Per esempio, se i dati storici terminano il 12/99 e si desidera iniziare le previsioni il 01/00, utilizzare il valore 1; se invece si desidera iniziare le previsioni il 03/00, utilizzare il valore 3.
- **Avvia da: Previsione storica.** Il periodo da cui si desidera far iniziare le previsioni storiche, espresso come distanza negativa dall'ultimo periodo dei dati storici di cui si dispone. Per esempio, se i dati storici terminano il 12/99 e si desidera fare previsioni storiche per gli ultimi cinque periodi di tempo dei propri dati, utilizzare il valore -5.

Termina stima. Specificare il periodo in cui si desidera terminare le previsioni.

- **Termina fase della previsione.** Il periodo in cui si desidera terminare le previsioni, espresso come distanza dall'ultimo periodo dei dati storici di cui si dispone. Per esempio, se i dati storici terminano il 12/99 e si desidera che le previsioni terminino il 6/00, utilizzare il valore 6. Per le previsioni future, il valore deve sempre essere superiore o uguale al valore Avvia da.

Nodo Cluster di sequenze MS

Il nodo Cluster di sequenze MS utilizza un algoritmo di analisi delle sequenze che esplora i dati contenenti eventi collegabili seguendo dei percorsi, o *sequenze*. Alcuni esempi potrebbero essere i percorsi di navigazione creati dagli utenti che consultano un sito Web, o l'ordine con cui un cliente

aggiunge gli articoli al carrello degli acquisti di un rivenditore online. L'algoritmo individua le sequenze più comuni raggruppando o *inserendo in cluster* le sequenze identiche.

Requisiti

I requisiti di un modello Cluster di sequenze Microsoft sono i seguenti:

- **Campo ID.** L'algoritmo di Cluster di sequenze Microsoft richiede che le informazioni delle sequenze siano memorizzate in formato transazionale (vedere Dati tabulari e dati transazionali a pag.). A tale fine è necessario disporre di un campo ID che identifichi le singole transazioni.
- **Almeno un campo di input.** L'algoritmo richiede almeno un campo di input.
- **Campo sequenza.** L'algoritmo richiede anche un campo di identificazione della sequenza che deve avere un livello di misurazione Continuo. Per esempio si può utilizzare un identificativo di pagina Web, un numero intero o una stringa di testo, purché il campo identifichi gli eventi di una sequenza. Per ogni sequenza è consentito un solo identificativo e per ogni modello è consentito un solo tipo di sequenza. Il campo Sequenza deve essere diverso dal campo ID e Univoco.
- **Campo obiettivo.** Quando si crea un modello di raggruppamento sequenze è necessario un campo obiettivo.
- **Campo univoco.** Un modello di cluster di sequenze richiede un campo chiave che identifichi i record in modo univoco. Il campo Univoco si può impostare sullo stesso valore del campo ID.

Opzioni dei campi Cluster di sequenze MS

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

Figura 3-16
Specifica dei campi per il cluster di sequenze MS

Per poter generare un modello di cluster di sequenze, è necessario prima specificare i campi da utilizzare come obiettivi e come input. Si noti che per il nodo Cluster di sequenze MS non è possibile utilizzare le informazioni dei campi di un nodo Tipo a monte: le impostazioni dei campi devono essere definite qui.

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).

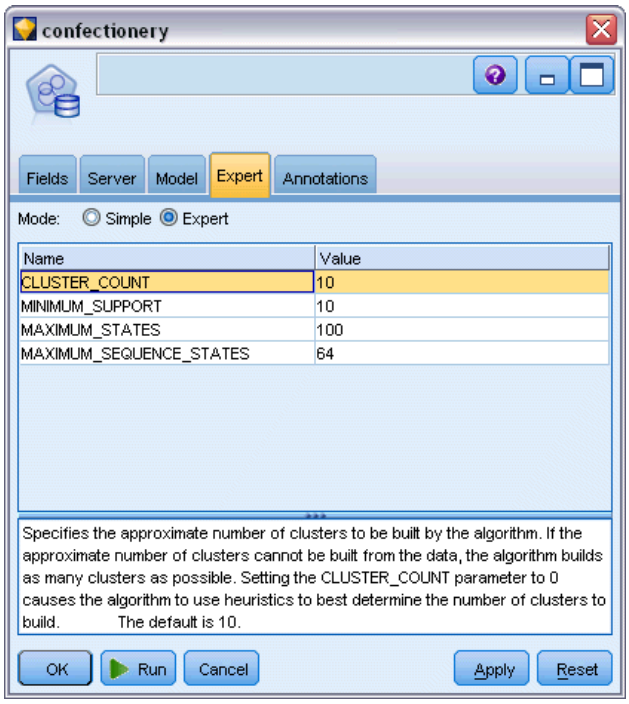
Input. Selezionare il campo o i campi di input per il modello. Questi sono i campi che contengono gli eventi rilevanti nella creazione di modelli di sequenza.

Sequenza. Scegliere dall'elenco un campo da utilizzare come campo identificativo della sequenza. Per esempio si può utilizzare un identificativo di pagina Web, un numero intero o una stringa di testo, purché il campo identifichi gli eventi di una sequenza. Per ogni sequenza è consentito un solo identificativo e per ogni modello è consentito un solo tipo di sequenza. Il campo Sequenza deve essere diverso dal campo ID (specificato in questa scheda) e dal campo Univoco (specificato nella scheda Modello).

Obiettivo. Scegliere un campo da utilizzare come campo obiettivo, cioè il campo di cui si sta cercando di prevedere il valore in base ai dati della sequenza.

Opzioni avanzate Cluster di sequenze MS

Figura 3-17
Specifica delle opzioni avanzate per il cluster di sequenze MS



Le opzioni disponibili nella scheda Livello avanzato possono variare a seconda della struttura dello stream selezionato. Per ulteriori dettagli sulle opzioni di livello avanzato disponibili per il nodo del modello Analysis Services selezionato, fare riferimento alla guida a livello di campo dell'interfaccia utente.

Calcolo del punteggio per i modelli di Analysis Services

Il calcolo del punteggio avviene in SQL Server ed è eseguito da Analysis Services. Può essere necessario caricare l'insieme di dati in una tabella temporanea, qualora i dati vengano originati in IBM® SPSS® Modeler o debbano essere preparati all'interno dell'applicazione. I modelli che l'utente crea da SPSS Modeler mediante il mining in-database rappresentano in effetti modelli remoti memorizzati nel server di database o di data mining. Si tratta di una distinzione importante da comprendere per la visualizzazione e il calcolo del punteggio dei modelli creati utilizzando gli algoritmi di Microsoft Analysis Services.

In SPSS Modeler viene generalmente fornita una sola previsione con la probabilità o la confidenza associata.

Per alcuni esempi di calcolo dei punteggi dei modelli, vedere Esempi di mining con Analysis Services a pag. 46.

Impostazioni comuni a tutti i modelli di Analysis Services

Le seguenti impostazioni sono valide per tutti i modelli di Analysis Services.

Scheda Server dell'insieme di modelli Analysis Services

La scheda Server consente di specificare le connessioni per il mining in-database. La scheda fornisce anche la chiave di modello univoca. Tale chiave è generata in modo casuale quando il modello viene creato e archiviato sia all'interno del modello di IBM® SPSS® Modeler che nella descrizione dell'oggetto modello memorizzata nel database di Analysis Services.

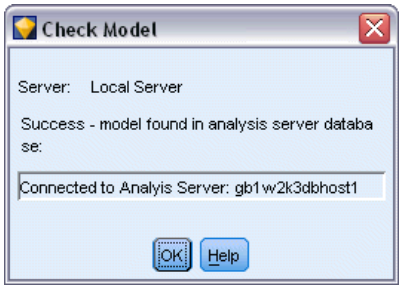
Figura 3-18
Opzioni Server per l'insieme di modelli Albero decisionale MS

Nella scheda Server è possibile configurare l'host e il database di Analysis Server nonché la sorgente dati SQL Server per le operazioni di calcolo del punteggio. Le opzioni specificate all'interno di questa scheda sovrascrivono quelle selezionate nelle finestre di dialogo Applicazioni di supporto o Creazione modello di IBM® SPSS® Modeler. Per ulteriori informazioni, vedere l'argomento Attivazione dell'integrazione con Analysis Services a pag. 17.

GUID modello. La chiave di modello viene visualizzata qui. Tale chiave è generata in modo casuale quando il modello viene creato e archiviato sia all'interno del modello di SPSS Modeler che nella descrizione dell'oggetto modello memorizzata nel database di Analysis Services.

Controllo. Fare clic su questo pulsante per confrontare la chiave qui visualizzata con quella all'interno del modello archiviato nel database di Analysis Services. Ciò consente di verificare se il modello è ancora presente in Analysis Server e se la relativa struttura non è stata modificata.

Figura 3-19
Risultati del controllo delle chiavi di modello

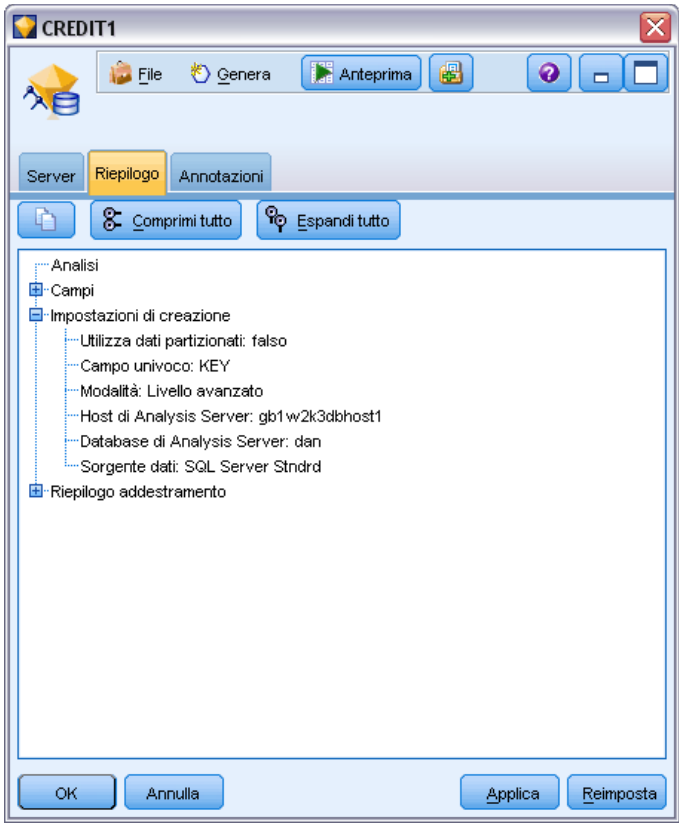


Nota: il pulsante Controllo è disponibile solo per i modelli aggiunti all'area di disegno dello stream per il successivo calcolo del punteggio. Qualora il controllo abbia esito negativo, verificare se il modello è stato eliminato o sostituito da un modello diverso sul server.

Visualizza. Fare clic per ottenere una visualizzazione grafica del modello di albero decisionale. La scheda Visualizzatore degli alberi decisionali è condivisa da altri algoritmi per alberi decisionali disponibili in SPSS Modeler e la funzionalità è identica. Per ulteriori informazioni, vedere l'argomento Insiemi di modelli Albero decisionale in il capitolo 6 in *IBM SPSS Modeler 15 Nodi Modelli*.

Scheda Riepilogo dell'insieme di modelli Analysis Services

Figura 3-20
Opzioni Riepilogo per l'insieme di modelli Albero decisionale MS



La scheda Riepilogo di un insieme di modelli 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 di espansione 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, utilizzare il controllo di espansione di un elemento se si desidera nascondere solo i risultati di tale elemento oppure fare clic sul pulsante Comprimi tutto se si desidera nascondere tutti i risultati.

Analisi. Visualizza informazioni sul modello specifico. Se è stato eseguito un nodo Analisi collegato a questo insieme di modelli, in questa sezione verranno visualizzate anche le informazioni sui risultati di tale analisi. Per ulteriori informazioni, vedere l'argomento nodo Analisi in il capitolo 6 in *IBM SPSS Modeler 15 Nodi di input, elaborazione e output*.

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, lo stream utilizzato per creare tale modello, l'utente che lo ha creato, la data di creazione e il tempo trascorso per la creazione del modello.

Insieme di modelli Serie storica MS

Il modello Serie storica MS genera punteggi solo per i periodi di tempo previsti, non per i dati storici.

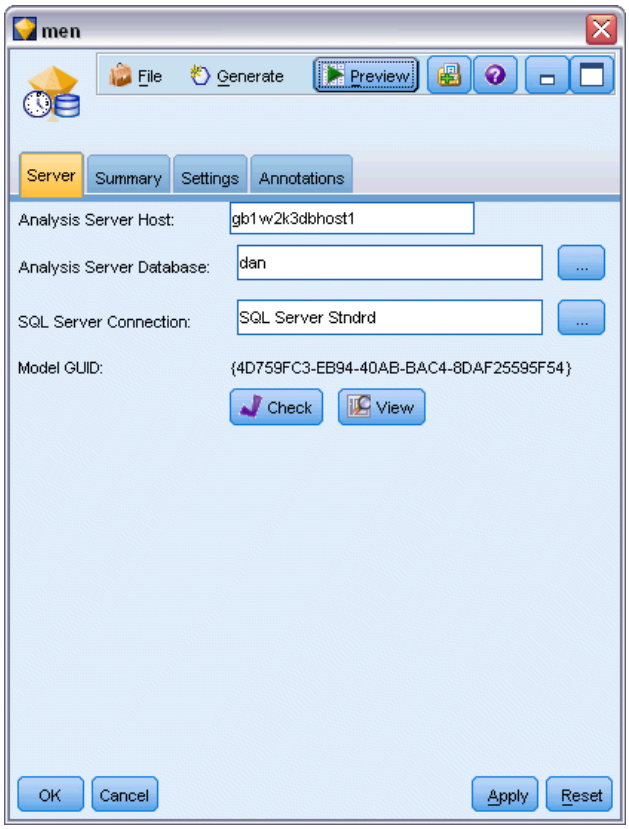
Al modello vengono aggiunti i seguenti campi:

Nome campo	**Descrizione**
$M-*campo*	Valore previsto del *campo*.
$Var-*campo*	Varianza calcolata del *campo*
$Stdev-*campo*	Deviazione standard del *campo*

Scheda Server dell'insieme di modelli Serie storica MS

La scheda Server consente di specificare le connessioni per il mining in-database. La scheda fornisce anche la chiave di modello univoca. Tale chiave è generata in modo casuale quando il modello viene creato e archiviato sia all'interno del modello di IBM® SPSS® Modeler che nella descrizione dell'oggetto modello memorizzata nel database di Analysis Services.

Figura 3-21
Opzioni della scheda Server per l'insieme di modelli Serie storica MS

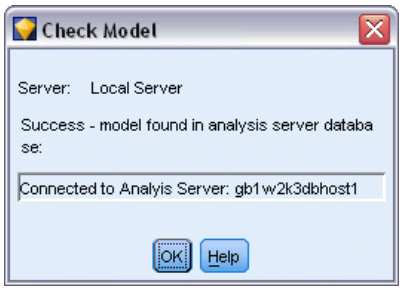


Nella scheda Server è possibile configurare l'host e il database di Analysis Server nonché la sorgente dati SQL Server per le operazioni di calcolo del punteggio. Le opzioni specificate all'interno di questa scheda sovrascrivono quelle selezionate nelle finestre di dialogo Applicazioni di supporto o Creazione modello di IBM® SPSS® Modeler. Per ulteriori informazioni, vedere l'argomento Attivazione dell'integrazione con Analysis Services a pag. 17.

GUID modello. La chiave di modello viene visualizzata qui. Tale chiave è generata in modo casuale quando il modello viene creato e archiviato sia all'interno del modello di SPSS Modeler che nella descrizione dell'oggetto modello memorizzata nel database di Analysis Services.

Controllo. Fare clic su questo pulsante per confrontare la chiave qui visualizzata con quella all'interno del modello archiviato nel database di Analysis Services. Ciò consente di verificare se il modello è ancora presente in Analysis Server e se la relativa struttura non è stata modificata.

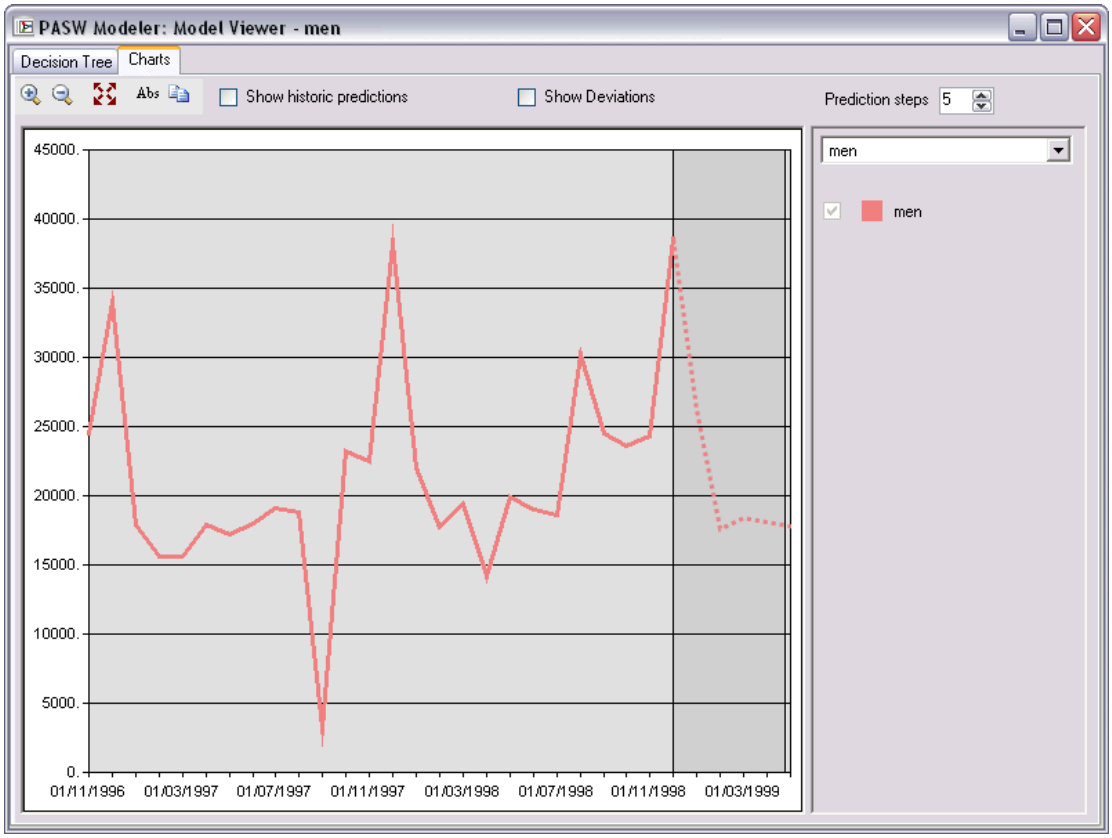
Figura 3-22
Risultati del controllo delle chiavi di modello



Nota: il pulsante Controllo è disponibile solo per i modelli aggiunti all'area di disegno dello stream per il successivo calcolo del punteggio. Qualora il controllo abbia esito negativo, verificare se il modello è stato eliminato o sostituito da un modello diverso sul server.

Visualizza. Fare clic per ottenere una rappresentazione grafica del modello di serie storica. Analysis Services mostra il modello completo sotto forma di albero. È possibile anche visualizzare un grafico che mostra il valore storico del campo obiettivo nel tempo, unitamente ai valori futuri previsti.

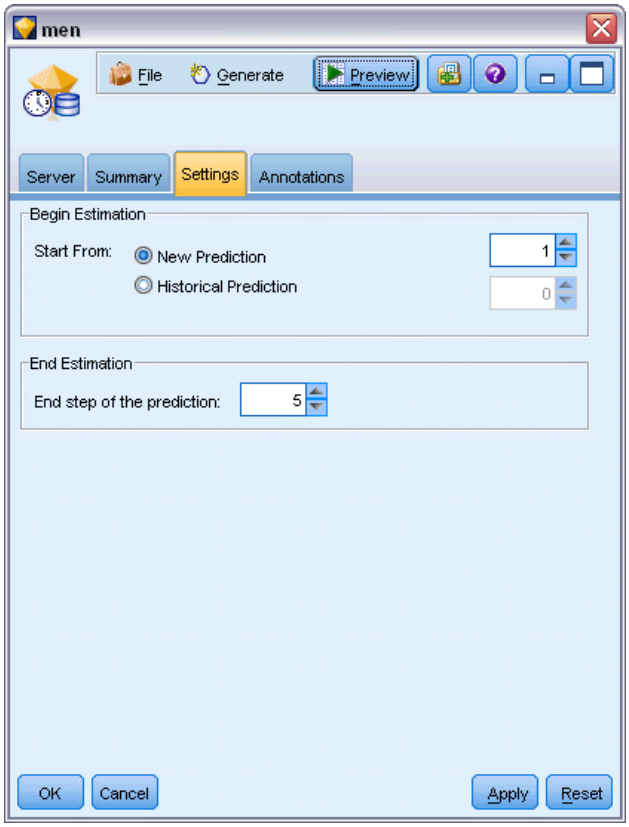
Figura 3-23
Visualizzatore serie storiche MS con i valori storici (linea continua) e i valori futuri previsti (linea tratteggiata)



Per ulteriori informazioni, vedere la descrizione del visualizzatore di serie storiche nella libreria MSDN all'indirizzo *http://msdn.microsoft.com/en-us/library/ms175331.aspx*.

Scheda Impostazioni dell'insieme di modelli Serie storica MS

Figura 3-24
Opzioni della scheda Impostazioni per l'insieme di modelli Serie storica MS



Inizia stima. Specificare il periodo da cui si desidera iniziare le previsioni.

- **Avvia da: Nuova previsione.** Il periodo da cui si desidera far iniziare le previsioni future, espresso come distanza dall'ultimo periodo dei dati storici di cui si dispone. Per esempio, se i dati storici terminano il 12/99 e si desidera iniziare le previsioni il 01/00, utilizzare il valore 1; se invece si desidera iniziare le previsioni il 03/00, utilizzare il valore 3.
- **Avvia da: Previsione storica.** Il periodo da cui si desidera far iniziare le previsioni storiche, espresso come distanza negativa dall'ultimo periodo dei dati storici di cui si dispone. Per esempio, se i dati storici terminano il 12/99 e si desidera fare previsioni storiche per gli ultimi cinque periodi di tempo dei propri dati, utilizzare il valore -5.

Termina stima. Specificare il periodo in cui si desidera terminare le previsioni.

- **Termina fase della previsione.** Il periodo in cui si desidera terminare le previsioni, espresso come distanza dall'ultimo periodo dei dati storici di cui si dispone. Per esempio, se i dati storici terminano il 12/99 e si desidera che le previsioni terminino il 6/00, utilizzare il valore 6. Per le previsioni future, il valore deve sempre essere superiore o uguale al valore Avvia da.

Insiemi di modelli Cluster di sequenze MS

Al modello Cluster di sequenze MS vengono aggiunti i seguenti campi (dove *campo* è il nome del campo obiettivo):

Nome campo	Descrizione
$MC-*campo*	Previsione del cluster a cui appartiene la sequenza.
$MCP-*campo*	Probabilità che la sequenza appartenga al cluster previsto.
$MS-*campo*	Valore previsto del *campo*.
$MSP-*campo*	Probabilità che il valore $MS-*campo* sia corretto.

Esportazione di modelli e generazione di nodi

È possibile esportare il riepilogo e la struttura di un modello in file formato testo e HTML, nonché generare i nodi Seleziona e Filtro appropriati laddove necessario. Per ulteriori informazioni, vedere l'argomento Esplorazione degli insiemi di modelli in il capitolo 3 in *IBM SPSS Modeler 15 Nodi Modelli*.

Analogamente ad altri insiemi di modelli in IBM® SPSS® Modeler, gli insiemi di modelli di Microsoft Analysis Services supportano la generazione diretta di nodi Operazioni su campi e record. Utilizzando le opzioni del menu Genera dell'insieme di modelli, è possibile generare i seguenti nodi:

- Nodo Seleziona (solo se è stato selezionato un elemento nella scheda Modello)
- nodo Filtro

Esempi di mining con Analysis Services

È disponibile un'ampia gamma di stream di esempio che illustrano l'utilizzo del data mining di MS Analysis Services con IBM® SPSS® Modeler. Tali stream si trovano nella cartella di installazione di SPSS Modeler in:

\Demos\Database_Modelling\Microsoft

Nota: alla cartella Demos è possibile accedere dal gruppo di programmi IBM SPSS Modeler del menu Start di Windows.

Stream di esempio: Alberi decisionali

Gli stream riportati di seguito possono essere utilizzati insieme, in ordine sequenziale, come esempio del processo di mining in-database basato sull'algoritmo per alberi decisionali fornito da MS Analysis Services.

Stream	Descrizione
1_upload_data.str	Utilizzato per la pulitura e il caricamento di dati da un file piatto nel database.
2_explore_data.str	Utilizzato come esempio di esplorazione dei dati con IBM® SPSS® Modeler.

Stream	Descrizione
3_build_model.str	Genera il modello utilizzando l'algoritmo nativo del database.
4_evaluate_model.str	Utilizzato come esempio di valutazione di modelli con SPSS Modeler.
5_deploy_model.str	Esegue il deployment del modello ai fini del calcolo del punteggio in-database.

Nota: Per eseguire l'esempio, gli stream devono essere eseguiti in ordine. Inoltre, i nodi Origine e Modelli in ogni stream devono essere aggiornati per far riferimento a un'origine dati valida per il database che si desidera utilizzare.

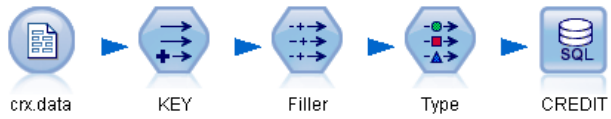
L'insieme di dati impiegato negli stream di esempio concerne applicazioni relative alle carte di credito e presenta un problema di classificazione relativo a un insieme misto di predittori continui e categoriali. Per ulteriori informazioni su questo insieme di dati, vedere il file *crx.names* nella stessa cartella degli stream di esempio.

Questo insieme di dati è disponibile in UCI Machine Learning Repository alla pagina *ftp://ftp.ics.uci.edu/pub/machine-learning-databases/credit-screening/*

Stream di esempio: Caricamento dati

Il primo stream di esempio, *1_upload_data.str*, viene utilizzato per pulire e caricare dati da un file piatto in SQL Server.

Figura 3-25
Stream di esempio usato per il caricamento dei dati



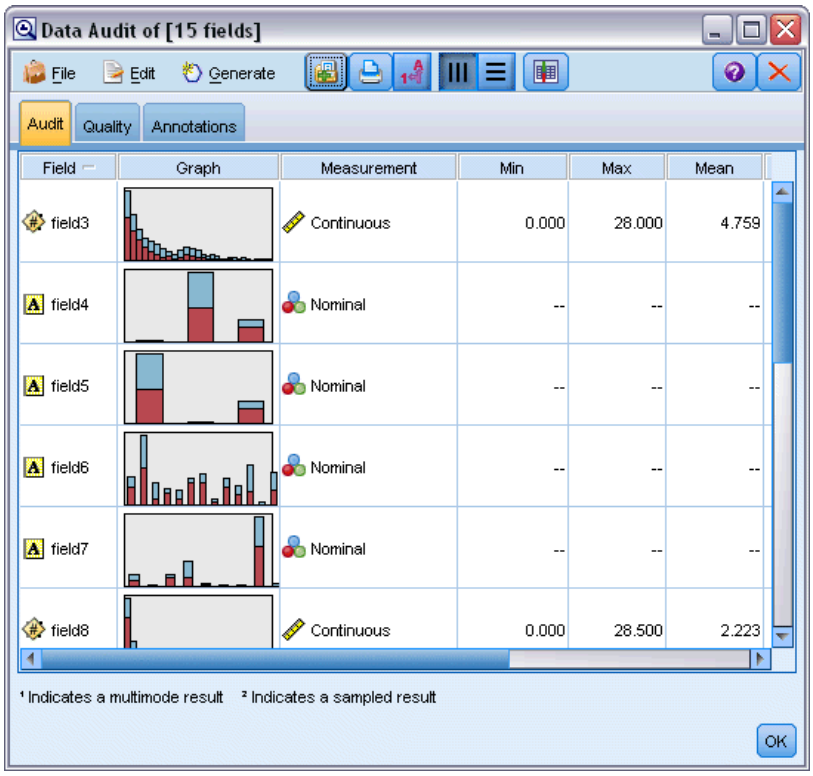
Poiché il data mining Analysis Services richiede la specifica di un campo chiave, questo stream iniziale utilizza un nodo Nuovo campo per aggiungere un nuovo campo all'insieme di dati denominato *CHIAVE* con i valori univoci *1,2,3* mediante la funzione @INDEX di IBM® SPSS® Modeler.

Il successivo nodo Riempimento viene utilizzato per la gestione dei valori mancanti e sostituisce i campi vuoti letti dal file di testo *crx.data* con valori *NULLO*.

Stream di esempio: Explore Data

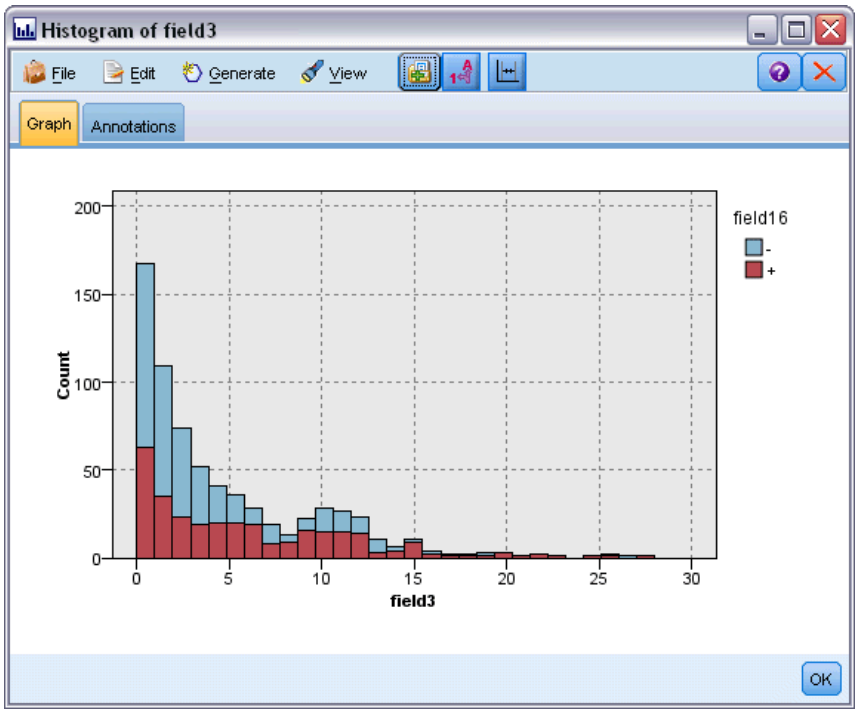
Il secondo stream di esempio, *2_explore_data.str*, viene utilizzato per illustrare l'uso di un nodo Esplora per acquisire una panoramica generale dei dati, comprese statistiche riassuntive e grafici. Per ulteriori informazioni, vedere l'argomento Nodo Esplora in il capitolo 6 in *IBM SPSS Modeler 15 Nodi di input, elaborazione e output*.

Figura 3-26
Risultati di Esplora



Facendo doppio clic su un grafico nel report Esplora, si accede a una visualizzazione più dettagliata del grafico che consente un'esplorazione più approfondita di un dato campo.

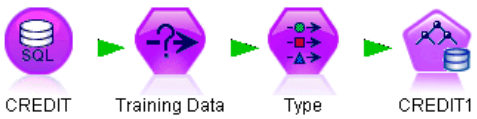
Figura 3-27
Istogramma creato facendo doppio clic su un grafico nella finestra Data Audit



Stream di esempio: Build Model

Il terzo stream di esempio, *3_build_model.str*, illustra la creazione di modelli in IBM® SPSS® Modeler. È possibile collegare il modello di database allo stream e fare doppio clic per specificare le impostazioni di creazione.

Figura 3-28
Stream di esempio relativo alla modellazione di database, in cui i nodi con ombreggiatura viola indicano l'esecuzione in-database



Nella scheda Modello della finestra di dialogo è possibile specificare quanto segue:

- Selezionare il campo Chiave come campo ID univoco.

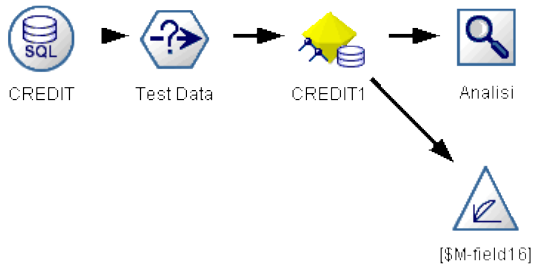
Nella scheda Livello avanzato è possibile regolare le impostazioni per la creazione del modello.

Prima di procedere all'esecuzione, assicurarsi di aver specificato il database corretto per la creazione del modello. Utilizzare la scheda Server per modificare le impostazioni.

Stream di esempio: Valutazione modello

Il quarto stream di esempio, *4_evaluate_model.str*, illustra i vantaggi associati all'utilizzo di IBM® SPSS® Modeler per la modellazione in-database. Una volta eseguito il modello, è possibile aggiungerlo nuovamente allo stream di dati e valutarlo con il supporto di un'ampia gamma di strumenti mirati disponibili in SPSS Modeler.

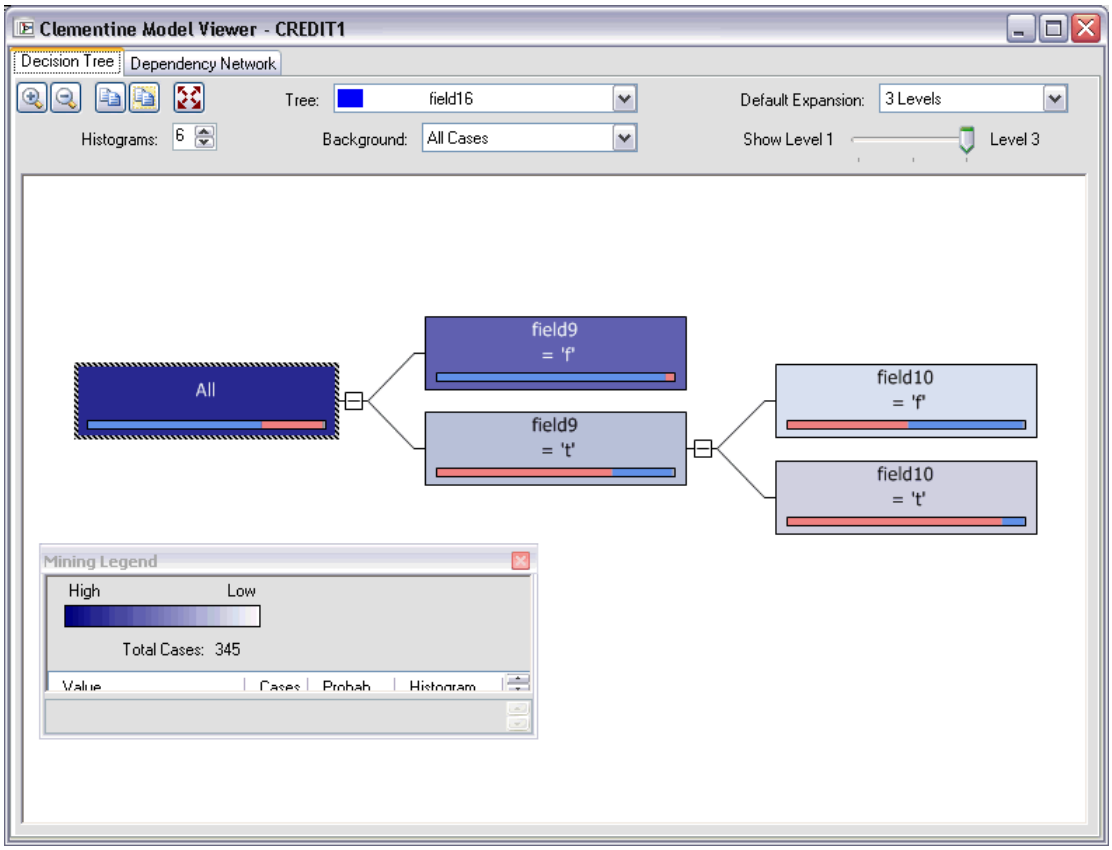
Figura 3-29
Stream di esempio usato per la valutazione del modello



Visualizzazione dei risultati della modellazione

È possibile fare doppio clic sull'insieme di modelli per esplorare i risultati. La scheda Riepilogo fornisce una visualizzazione dei risultati con struttura ad albero di regole. È inoltre possibile fare clic sul pulsante Visualizza della scheda Server per visualizzare graficamente il modello Alberi decisionali.

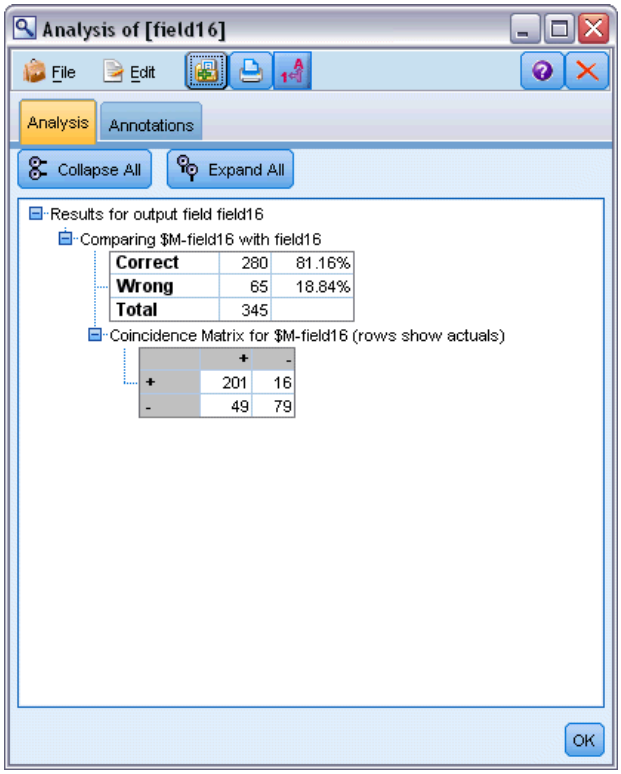
Figura 3-30
Visualizzatore che fornisce una rappresentazione grafica dei risultati della modellazione con albero decisionale MS



Valutazione dei risultati della modellazione

Il nodo Analisi nello stream di esempio crea una matrice di coincidenza che mostra lo schema di corrispondenze tra ogni campo previsto e il relativo campo obiettivo. Eseguire il nodo Analisi per visualizzare i risultati.

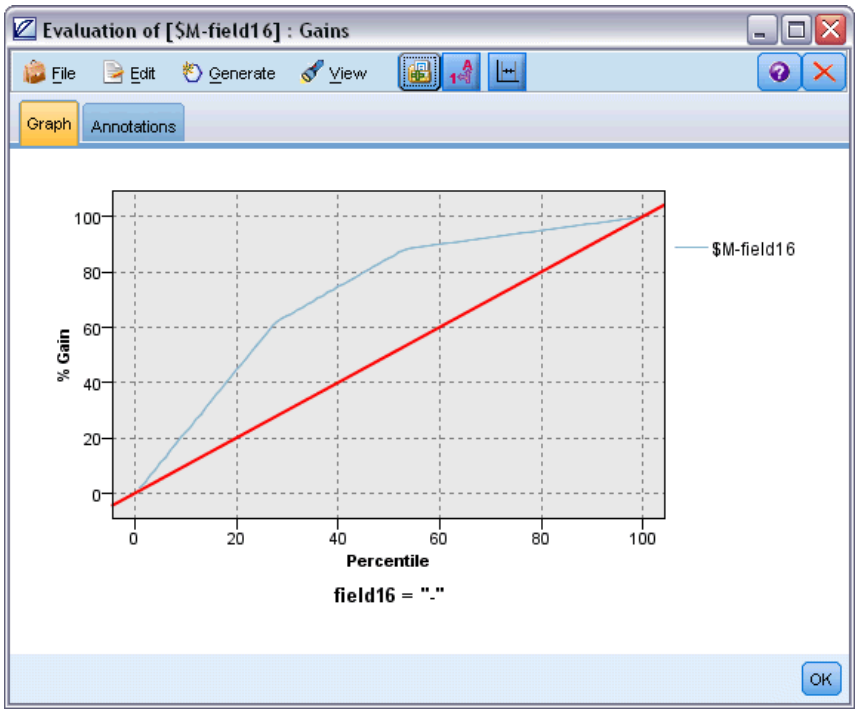
Figura 3-31
Risultati del nodo Analisi



La tabella indica che l'81,16% delle previsioni generate dall'algoritmo Albero decisionale MS era corretto.

Il nodo Valutazione nello stream di esempio può creare un grafico dei guadagni, progettato per mostrare i miglioramenti in termini di precisione realizzati dal modello. Eseguire il nodo Valutazione per visualizzare i risultati.

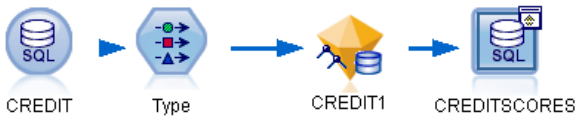
Figura 3-32
Grafico dei guadagni generato mediante il nodo Valutazione



Stream di esempio: Deployment modello

Una volta raggiunto il livello di precisione del modello desiderato, è possibile eseguire il deployment del modello per consentirne l'utilizzo con applicazioni esterne o la ripubblicazione nel database. Nell'ultimo stream di esempio, *5_deploy_model.str*, i dati vengono letti dalla tabella CREDIT, quindi viene eseguito il calcolo del punteggio e, infine, i dati vengono pubblicati nella tabella CREDITSCORES mediante il nodo di esportazione Database.

Figura 3-33
Stream di esempio usato per il deployment del modello



L'esecuzione dello stream genera il seguente codice SQL:

```
DROP TABLE CREDITSCORES

CREATE TABLE CREDITSCORES ( "field1" varchar(1),"field2" varchar(255),"field3" f
loat,"field4" varchar(1),"field5" varchar(2),"field6" varchar(2),"field7" varcha
r(2),"field8" float,"field9" varchar(1),"field10" varchar(1),"field11" int,"fiel
d12" varchar(1),"field13" varchar(1),"field14" int,"field15" int,"field16" varch
ar(1),"KEY" int,"$M-field16" varchar(9),"$MC-field16" float )
```

```
INSERT INTO CREDITSCORES ("field1","field2","field3","field4","field5","field6","field7","field8",
"field9","field10","field11","field12","field13","field14","field15","field16",
"KEY","$M-field16","$MC-field16")
SELECT T0.C0 AS C0,T0.C1 AS C1,T0.C2 AS C2,T0.C3 AS C3,T0.C4 AS C4,T0.C5 AS C5,
   T0.C6 AS C6,T0.C7 AS C7,T0.C8 AS C8,T0.C9 AS C9,T0.C10 AS C10,
   T0.C11 AS C11,T0.C12 AS C12,T0.C13 AS C13,T0.C14 AS C14,
   T0.C15 AS C15,T0.C16 AS C16,T0.C17 AS C17,T0.C18 AS C18
FROM (
   SELECT CONVERT(NVARCHAR,[TA].[field1]) AS C0, CONVERT(NVARCHAR,[TA].[field2]) AS C1,
      [TA].[field3] AS C2, CONVERT(NVARCHAR,[TA].[field4]) AS C3,
      CONVERT(NVARCHAR,[TA].[field5]) AS C4, CONVERT(NVARCHAR,[TA].[field6]) AS C5,
      CONVERT(NVARCHAR,[TA].[field7]) AS C6, [TA].[field8] AS C7,
      CONVERT(NVARCHAR,[TA].[field9]) AS C8, CONVERT(NVARCHAR,[TA].[field10]) AS C9,
      [TA].[field11] AS C10, CONVERT(NVARCHAR,[TA].[field12]) AS C11,
      CONVERT(NVARCHAR,[TA].[field13]) AS C12, [TA].[field14] AS C13,
      [TA].[field15] AS C14, CONVERT(NVARCHAR,[TA].[field16]) AS C15,
      [TA].[KEY] AS C16, CONVERT(NVARCHAR,[TA].[$M-field16]) AS C17,
      [TA].[$MC-field16] AS C18
   FROM openrowset('MSOLAP',
      'Datasource=localhost;Initial catalog=FoodMart 2000',
      'SELECT [T].[C0] AS [field1],[T].[C1] AS [field2],[T].[C2] AS [field3],
         [T].[C3] AS [field4],[T].[C4] AS [field5],[T].[C5] AS [field6],
         [T].[C6] AS [field7],[T].[C7] AS [field8],[T].[C8] AS [field9],
         [T].[C9] AS [field10],[T].[C10] AS [field11],[T].[C11] AS [field12],
         [T].[C12] AS [field13],[T].[C13] AS [field14],[T].[C14] AS [field15],
         [T].[C15] AS [field16],[T].[C16] AS [KEY],[CREDIT1].[field16] AS [$M-field16],
         PredictProbability([CREDIT1].[field16]) AS [$MC-field16]
      FROM [CREDIT1] PREDICTION JOIN
         openrowset(''MSDASQL'',
         ''Dsn=LocalServer;Uid=;pwd='',''SELECT T0."field1" AS C0,T0."field2" AS C1,
         T0."field3" AS C2,T0."field4" AS C3,T0."field5" AS C4,T0."field6" AS C5,
         T0."field7" AS C6,T0."field8" AS C7,T0."field9" AS C8,T0."field10" AS C9,
         T0."field11" AS C10,T0."field12" AS C11,T0."field13" AS C12,
         T0."field14" AS C13,T0."field15" AS C14,T0."field16" AS C15,
         T0."KEY" AS C16 FROM "dbo".CREDITDATA T0'') AS [T]
      ON [T].[C2] = [CREDIT1].[field3] and [T].[C7] = [CREDIT1].[field8]
         and [T].[C8] = [CREDIT1].[field9] and [T].[C9] = [CREDIT1].[field10]
         and [T].[C10] = [CREDIT1].[field11] and [T].[C11] = [CREDIT1].[field12]
         and [T].[C14] = [CREDIT1].[field15]') AS [TA]
) T0
```

Capitolo
4

Modellazione di database con Oracle Data Mining

Informazioni su Oracle Data Mining

IBM® SPSS® Modeler supporta l'integrazione con Oracle Data Mining (ODM), che include una serie di algoritmi di data mining saldamente incorporati nel sistema Oracle RDBMS. Queste funzionalità sono accessibili tramite l'interfaccia utente grafica e l'ambiente di sviluppo basato sui flussi di lavoro di SPSS Modeler e consentono ai clienti di sfruttare i vantaggi offerti dagli algoritmi di data mining ODM.

SPSS Modeler supporta l'integrazione dei seguenti algoritmi di Oracle Data Mining:

- Bayes naive
- Bayes adattivo
- Support Vector Machine (SVM)
- Modelli lineari generalizzati (GLM)*
- Albero decisionale
- O-Cluster
- K-Means
- NMF (fattorizzazione a matrice non negativa)
- Apriori
- MDL (Lunghezza descrizione minima)
- Importanza attributo (AI)

* 11*g* R1 soltanto

Requisiti per l'integrazione con Oracle

Di seguito sono elencate le condizioni che costituiscono i prerequisiti indispensabili per l'esecuzione della modellazione in-database con Oracle Data Mining. Per garantire che queste condizioni vengano soddisfatte, può essere necessario consultare l'amministratore di sistema.

- Esecuzione di IBM® SPSS® Modeler in modalità locale o in un'installazione di IBM® SPSS® Modeler Server su Windows o UNIX.
- Oracle 10*g*R2 o 11*g*R1 (Database 10.2 o versione successiva) con l'opzione Oracle Data Mining.

Nota: 10*g*R2 supporta tutti gli algoritmi di modellazione in-database ad eccezione di Modelli lineari generalizzati (richiede 11*g*R1).

- Una sorgente dati ODBC per la connessione a Oracle, come illustrato di seguito.

Nota: le funzionalità di modellazione in-database e ottimizzazione SQL richiedono che sul computer SPSS Modeler sia attivata la connettività SPSS Modeler Server. Con questa impostazione attivata, è possibile accedere agli algoritmi di database, restituire codice SQL direttamente da SPSS Modeler e accedere a SPSS Modeler Server. Per verificare lo stato attuale della licenza, scegliere le seguenti opzioni dal menu SPSS Modeler.
Guida > Informazioni su > Ulteriori dettagli

Se la connettività è abilitata, l'opzioneAbilitazione server viene visualizzata nella scheda Stato della licenza.

Per ulteriori informazioni, vedere l'argomento Connessione a IBM SPSS Modeler Server in il capitolo 3 in *Manuale dell'utente di IBM SPSS Modeler 15*.

Attivazione dell'integrazione con Oracle

Per attivare l'integrazione di IBM® SPSS® Modeler con Oracle Data Mining, sarà necessario configurare Oracle e creare una sorgente ODBC, attivare l'integrazione nella finestra di dialogo Applicazioni di supporto di SPSS Modeler e abilitare la generazione e l'ottimizzazione SQL.

Configurazione di Oracle

Per installare e configurare Oracle Data Mining, consultare la documentazione Oracle—in particolare la *Oracle Administrator's Guide*—.

Creazione di una sorgente ODBC per Oracle

Per attivare la connessione tra Oracle e SPSS Modeler è necessario creare un nome di sorgente dati (DSN, Data Source Name) ODBC.

Per creare un DSN, è necessario avere una conoscenza di base delle sorgenti dati e dei driver ODBC e disporre del supporto database in SPSS Modeler. Per ulteriori informazioni, vedere l'argomento Accesso ai dati in il capitolo 2 in *IBM SPSS Modeler Server 15 Guida della performance e amministrazione*.

Se l'applicazione è in esecuzione in modalità distribuita su IBM® SPSS® Modeler Server, creare il DSN sul computer server. Se invece è attiva la modalità locale (client), creare il DSN sul computer client.

- Installare i driver ODBC. I driver sono disponibili sul disco di installazione di IBM® SPSS® Data Access Pack fornito con questa versione. Eseguire il file *setup.exe* per avviare il programma di installazione, e selezionare tutti i driver opportuni. Attenersi alle istruzioni visualizzate per installare i driver.

- Creare il DSN.

  Nota: la sequenza dei menu dipende dalla versione di Windows in uso.

  - **Windows XP.** Dal menu Start, scegliere Pannello di controllo. Fare doppio clic su Strumenti di amministrazione, quindi ancora doppio clic su Origini dati (ODBC).

- **Windows Vista.** Dal menu Start, scegliere Pannello di controllo, quindi Strumenti di amministrazione. Fare doppio clic su Strumenti di amministrazione, selezionare Origini dati (ODBC) quindi Apri.
- **In Windows 7.** Dal menu Start, scegliere Pannello di controllo, quindi Sistema e sicurezza e Strumenti di amministrazione. Selezionare Origini dati (ODBC) e fare clic su Apri.

▶ Fare clic sulla scheda DSN di sistema, quindi fare clic su Aggiungi.

▶ Selezionare il driver SPSS OEM 6.0 Oracle Wire Protocol.

▶ Fare clic su Fine.

▶ Nella schermata di impostazione del driver ODBC Oracle Wire Protocol immettere il nome di una sorgente dati a scelta, il nome host del server Oracle, il numero di porta per la connessione e il SID dell'istanza Oracle in uso.

Nome host, numero di porta e SID possono essere ottenuti dal file *tnsnames.ora*, presente sul computer server, se è stato implementato TNS con un file *tnsnames.ora*. Per ulteriori informazioni, contattare l'amministratore Oracle.

▶ Fare clic sul pulsante Test per verificare la connessione.

Attivazione dell'integrazione di Oracle Data Mining in IBM SPSS Modeler

▶ Dai menu di SPSS Modeler scegliere:
Strumenti > Opzioni > Applicazioni di supporto

▶ Fare clic sulla scheda Oracle.

Attiva integrazione di Oracle Data Mining. Attiva la palette Modelli in-database (se non è già visualizzata) nella parte inferiore della finestra SPSS Modeler e aggiunge i nodi degli algoritmi di Oracle Data Mining.

Connessione Oracle. Specificare la sorgente dati ODBC Oracle di default, utilizzata per la creazione e l'archiviazione di modelli, insieme a un nome utente e una password validi. Questa impostazione può essere sovrascritta nei singoli nodi modelli e insiemi di modelli.

Nota: la connessione al database utilizzata a fini di modellazione può corrispondere o meno a quella impiegata per accedere ai dati. Per esempio, è possibile utilizzare uno stream che accede ai dati di un database Oracle, li scarica in SPSS Modeler per la pulitura o altre operazioni di modifica e, infine, li carica in un database Oracle differente per la modellazione. In alternativa, i dati originali possono risiedere in un file piatto o in un'altra sorgente (non Oracle), nel qual caso sarà necessario caricarli in Oracle per il processo di modellazione. In tutti i casi, i dati verranno automaticamente caricati in una tabella temporanea creata nel database utilizzato per la modellazione.

Avvisa prima di sovrascrivere un modello Oracle Data Mining. Selezionare questa opzione per assicurarsi che i modelli archiviati nel database non vengano sovrascritti da SPSS Modeler senza preavviso.

Elenca modelli Oracle Data Mining. Visualizza i modelli di data mining disponibili.

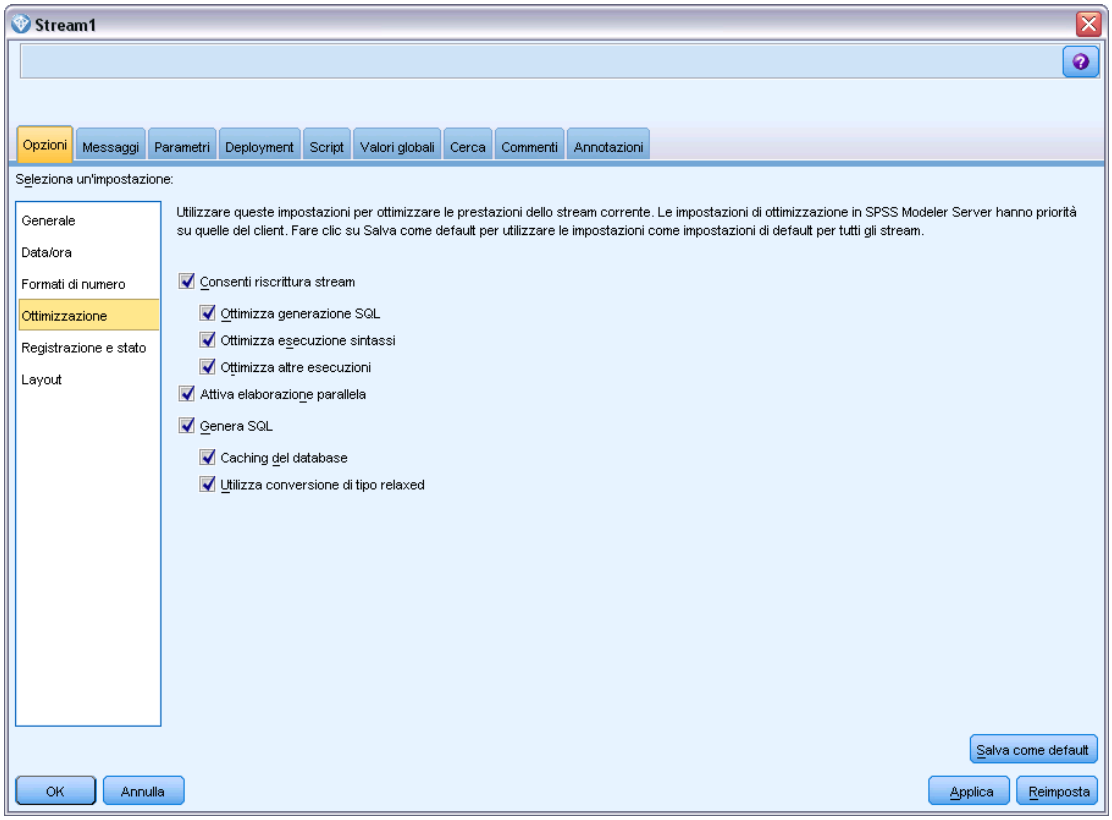
Attiva avvio di Oracle Data Miner. (facoltativo) Se attivata, consente a SPSS Modeler di avviare l'applicazione Oracle Data Miner. Per ulteriori informazioni, fare riferimento a Oracle Data Miner a pag. 96.

Percorso file eseguibile di Oracle Data Miner. (facoltativo) Specifica la posizione fisica del file eseguibile di Oracle Data Miner per Windows (per esempio *C:\odm\bin\odminerw.exe*). Oracle Data Miner non viene installato insieme a SPSS Modeler; è necessario scaricare la versione corretta dal sito Web di Oracle (*http://www.oracle.com/technology/products/bi/odm/odminer.html*) e installarla sul client.

Attivazione di generazione e ottimizzazione SQL

- Dai menu di SPSS Modeler scegliere:
  Strumenti > Proprietà stream > Opzioni

Figura 4-1
Impostazioni di ottimizzazione



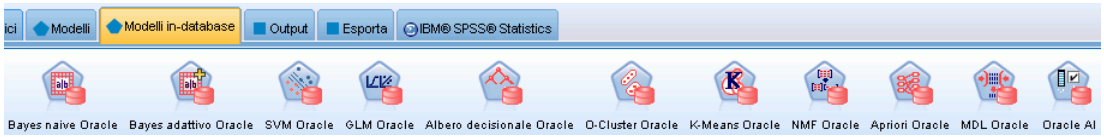
- Fare clic sull'opzione Ottimizzazione nel riquadro di spostamento.
- Confermare che l'opzione Genera SQL è attivata. Questa impostazione è necessaria per il corretto funzionamento della modellazione di database.
- Selezionare Ottimizza generazione SQL e Ottimizza altre esecuzioni (queste due opzioni non sono strettamente necessarie, tuttavia se ne consiglia la selezione per ottenere performance ottimizzate).

Per ulteriori informazioni, vedere l'argomento Impostazione delle opzioni di ottimizzazione per gli stream in il capitolo 5 in *Manuale dell'utente di IBM SPSS Modeler 15*.

Creazione di modelli con Oracle Data Mining

I nodi di creazione modelli Oracle funzionano esattamente come gli altri nodi Modelli di IBM® SPSS® Modeler con alcune eccezioni. È possibile accedere a questi nodi dalla palette Modelli in-database, presente nella parte inferiore della finestra di SPSS Modeler.

Figura 4-2
Palette Modelli in-database



Considerazioni sui dati

Oracle richiede che i dati categoriali siano archiviati in formato stringa (CHAR o VARCHAR2). Di conseguenza, SPSS Modeler non consentirà di specificare campi di archiviazione numerici con livello di misurazione *Flag* o *Nominale* (categoriali) come input per modelli ODM. Se necessario, i numeri possono essere convertiti in stringhe in SPSS Modeler utilizzando il nodo Ricodifica. Per ulteriori informazioni, vedere l'argomento Nodo Ricodifica in il capitolo 4 in *IBM SPSS Modeler 15 Nodi di input, elaborazione e output*.

Campo obiettivo. Nei modelli di classificazione ODM è possibile selezionare un solo campo come campo di output (obiettivo).

Nome modello. A partire da Oracle 11*g*R1, il nome unique è una parola chiave e non può essere utilizzato come nome di modello personalizzato.

Campo univoco. Specifica il campo utilizzato per identificare ciascun caso in modo univoco. Tale campo può coincidere, per esempio, con un campo ID quale *IDCliente*. SPSS Modeler impone che questo campo chiave debba essere numerico.

Nota: questo campo è opzionale per tutti i nodi Oracle, ad eccezione dei nodi Bayes adattivi, O-Cluster e Apriori Oracle.

Commenti generali

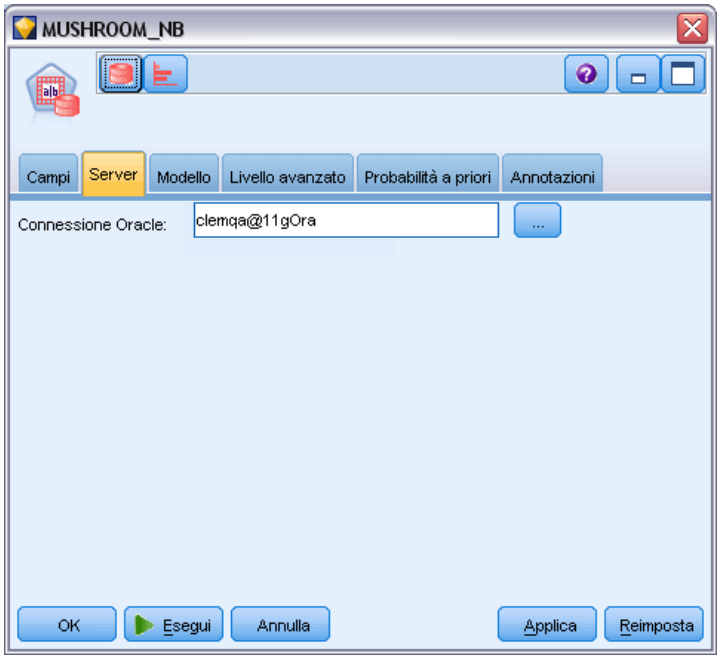
- SPSS Modeler non consente di eseguire operazioni di esportazione e importazione PMML per i modelli creati mediante Oracle Data Mining.
- In ODM viene sempre eseguito il calcolo del punteggio dei modelli. Può essere necessario caricare l'insieme di dati in una tabella temporanea, qualora i dati vengano originati in SPSS Modeler o debbano essere preparati all'interno dell'applicazione.
- In SPSS Modeler viene generalmente fornita una sola previsione con la probabilità o la confidenza associata.

- SPSS Modeler restringe a 1.000 il numero di campi utilizzabili per la creazione di modelli e il calcolo del punteggio.
- SPSS Modeler è in grado di calcolare il punteggio dei modelli ODM dall'interno di stream pubblicati per l'esecuzione utilizzando IBM® SPSS® Modeler Solution Publisher. Per ulteriori informazioni, vedere l'argomento Funzionamento di IBM SPSS Modeler Solution Publisher in il capitolo 2 in *IBM SPSS Modeler 15 Solution Publisher*.

Opzioni della scheda Server dei modelli Oracle

Specificare la connessione Oracle utilizzata per caricare i dati per la modellazione. Se necessario, inoltre, è possibile selezionare nella scheda Server una connessione specifica per ogni nodo Modelli che sovrascriva la connessione Oracle di default indicata nella finestra di dialogo Applicazioni di supporto. Per ulteriori informazioni, vedere l'argomento Attivazione dell'integrazione con Oracle a pag. 56.

Figura 4-3
Opzioni Server Oracle



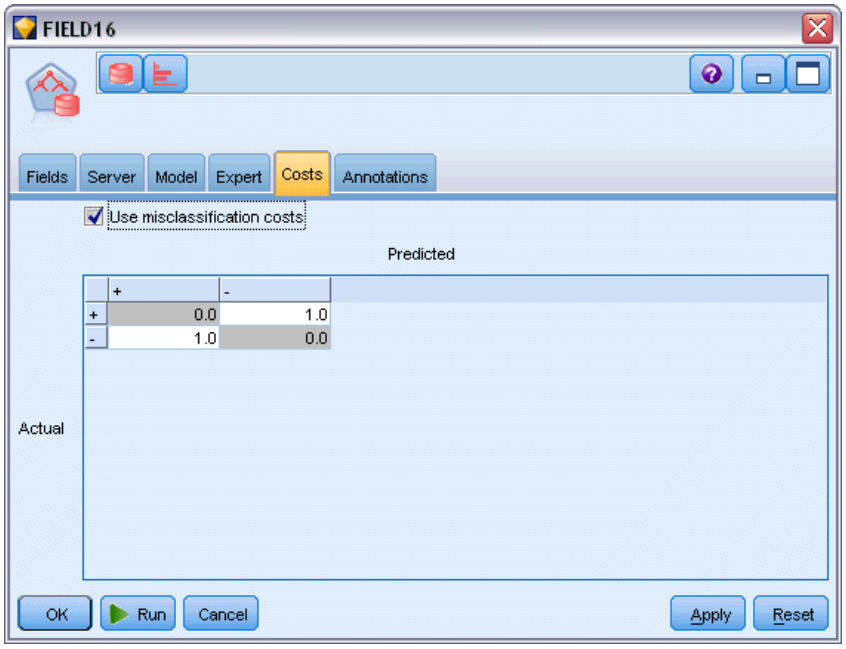
Commenti

- La connessione utilizzata per la modellazione può corrispondere o meno a quella impiegata nel nodo di input di uno stream. Per esempio, è possibile utilizzare uno stream che accede ai dati di un database Oracle, li scarica in IBM® SPSS® Modeler per la pulitura o altre operazioni di modifica e, infine, li carica in un database Oracle differente per la modellazione.
- Il nome della sorgente dati ODBC è efficacemente incorporato in ogni stream di SPSS Modeler. Se uno stream creato su un determinato host viene eseguito su un host differente, il nome della sorgente dati deve essere lo stesso su entrambi gli host. In alternativa, è possibile

selezionare una sorgente dati differente nella scheda Server all'interno di ogni nodo Modelli o di input.

Costi classificazione errata

Figura 4-4
Opzioni della scheda Costi Oracle



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.

Nota: solo il modello Alberi decisionali consente la specifica dei costi al momento della creazione.

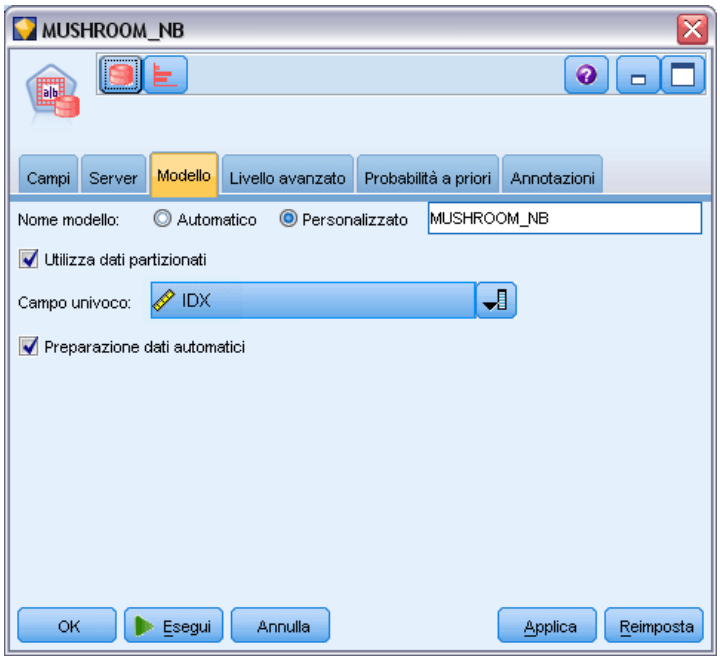
Bayes naive Oracle

Bayes naive è un algoritmo molto noto per problemi di classificazione. Il modello viene definito *naïve* perché considera tutte le variabili di previsione proposte come indipendenti l'una dall'altra. Bayes naive è un algoritmo veloce e scalabile in grado di calcolare probabilità condizionali per combinazioni di attributi e per l'attributo obiettivo. Dai dati di addestramento viene stabilita una probabilità indipendente che serve a indicare la verosimiglianza di ciascuna classe obiettivo una volta specificata l'occorrenza di ogni categoria di valore da ogni variabile di input.

- La validazione incrociata viene utilizzata per verificare la precisione di un modello con gli stessi dati adoperati per creare il modello. Questa operazione risulta particolarmente utile quando il numero di casi disponibili per creare un modello è ridotto.
- L'output del modello può essere visualizzato in formato matrice. I numeri della matrice indicano probabilità condizionali che mettono in relazione le classi previste (colonne) e le combinazioni di valori-variabili dei predittori (righe).

Opzioni del modello Bayes naive

Figura 4-5
Opzioni del modello Bayes naive



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 di partizione, questa opzione garantisce che per la creazione del modello verranno utilizzati solo i dati della partizione di addestramento. Per ulteriori informazioni, vedere l'argomento Nodo Partizione in il capitolo 4 in *IBM SPSS Modeler 15 Nodi di input, elaborazione e output*.

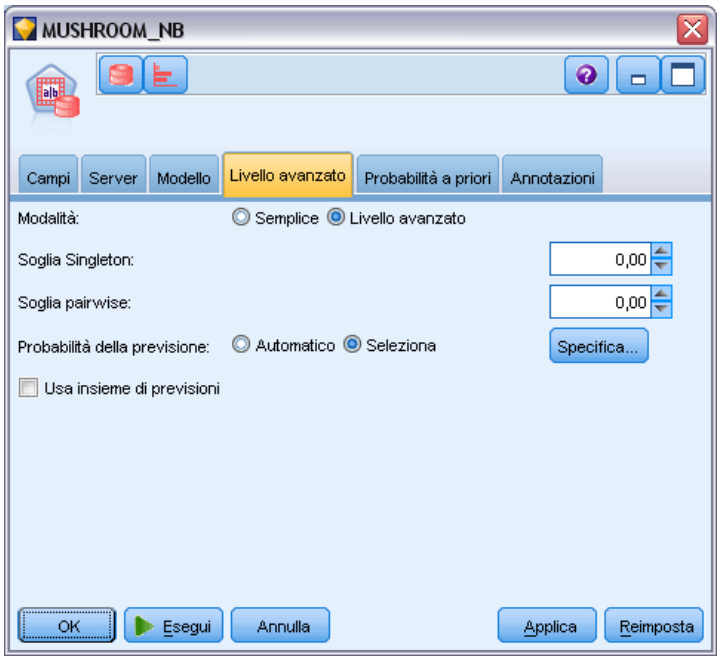
Campo univoco. Specifica il campo utilizzato per identificare ciascun caso in modo univoco. Tale campo può coincidere, per esempio, con un campo ID quale *IDCliente*. IBM® SPSS® Modeler impone che questo campo chiave debba essere numerico.

Nota: questo campo è opzionale per tutti i nodi Oracle, ad eccezione dei nodi Bayes adattivi, O-Cluster e Apriori Oracle.

Preparazione dati automatici. (solo 11*g*) Attiva (default) o disattiva la modalità di preparazione dei dati automatici di Oracle Data Mining. Se la casella è selezionata, ODM esegue automaticamente la trasformazione dei dati richiesti dall'algoritmo. Per ulteriori informazioni, vedere *Oracle Data Mining Concepts*.

Opzioni avanzate di Bayes naive

Figura 4-6
Opzioni avanzate di Bayes naive



Al momento della creazione del modello, i valori o le coppie di valori dei singoli attributi dei predittori vengono ignorati, a meno che non ci siano sufficienti occorrenze di un determinato valore o coppia di valori nei dati di addestramento. Le soglie per ignorare i valori vengono

specificate come frazioni basate sul numero di record presenti nei dati di addestramento. Adeguando queste soglie è possibile ridurre il rumore e migliorare la capacità del modello di essere generalizzato per altri insiemi di dati.

- **Soglia Singleton.** Specifica la soglia per un determinato valore di attributo predittore. Il numero di occorrenze di un determinato valore deve essere uguale o maggiore della frazione specificata, altrimenti il valore verrà ignorato.
- **Soglia pairwise.** Specifica la soglia per una determinata coppia di valori di attributo e predittore. Il numero di occorrenze di una determinata coppia di valori deve essere uguale o maggiore della frazione specificata, altrimenti la coppia verrà ignorata.

Probabilità della previsione. Consente al modello di includere la probabilità di una previsione corretta di un possibile risultato del campo obiettivo. Per abilitare questa funzione, scegliere Seleziona, fare clic sul pulsante Specifica, scegliere uno dei risultati possibili e fare clic su Inserisci.

Usa insieme di previsioni. Genera una tabella di tutti i risultati possibili relativi a tutti gli esiti possibili del campo obiettivo.

Bayes adattivi Oracle

La rete di Bayes adattivi (ABN) crea classificatori di rete bayesiana utilizzando la lunghezza di descrizione minima (MDL) e la selezione automatica delle funzionalità. ABN funziona bene in alcune situazioni in cui il modello Bayes naive non garantisce performance adeguate e in molti altri casi, sebbene con performance meno elevate. L'algoritmo ABN consente di creare tre tipi di modelli avanzati su base bayesiana, tra cui i modelli di albero decisionale semplificato (funzione singola), Bayes naive tagliato e multifunzione boosted.

Modelli generati

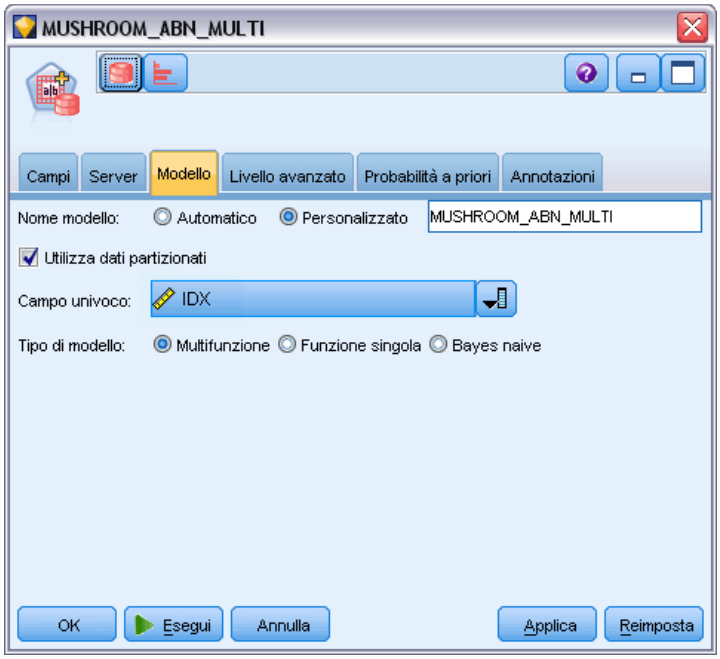
Nella modalità di creazione a funzione singola, ABN produce un albero decisionale semplificato basato su un insieme di regole leggibili, che consente all'utente aziendale o all'analista di comprendere la base delle previsioni del modello, per agire di conseguenza o fornire spiegazioni ad altri. Consente di ottenere un vantaggio significativo rispetto ai modelli Bayes naive e multifunzione. Queste regole possono essere visualizzate come un insieme di regole standard in IBM® SPSS® Modeler. Un semplice insieme di regole potrebbe avere il seguente aspetto:

IF MARITAL_STATUS = "Coniugato"
AND EDUCATION_NUM = "13-16"
THEN CHURN= "VERO"
Confidenza = .78, Supporto = 570 casi

I modelli Bayes naive tagliato e multifunzione non possono essere visualizzati in SPSS Modeler.

Opzioni del modello Bayes adattivo

Figura 4-7
Opzioni del modello Bayes adattivo



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 di partizione, questa opzione garantisce che per la creazione del modello verranno utilizzati solo i dati della partizione di addestramento. Per ulteriori informazioni, vedere l'argomento Nodo Partizione in il capitolo 4 in *IBM SPSS Modeler 15 Nodi di input, elaborazione e output*.

Campo univoco. Specifica il campo utilizzato per identificare ciascun caso in modo univoco. Tale campo può coincidere, per esempio, con un campo ID quale *IDCliente*. IBM® SPSS® Modeler impone che questo campo chiave debba essere numerico.

Nota: questo campo è opzionale per tutti i nodi Oracle, ad eccezione dei nodi Bayes adattivi, O-Cluster e Apriori Oracle.

Tipo di modello

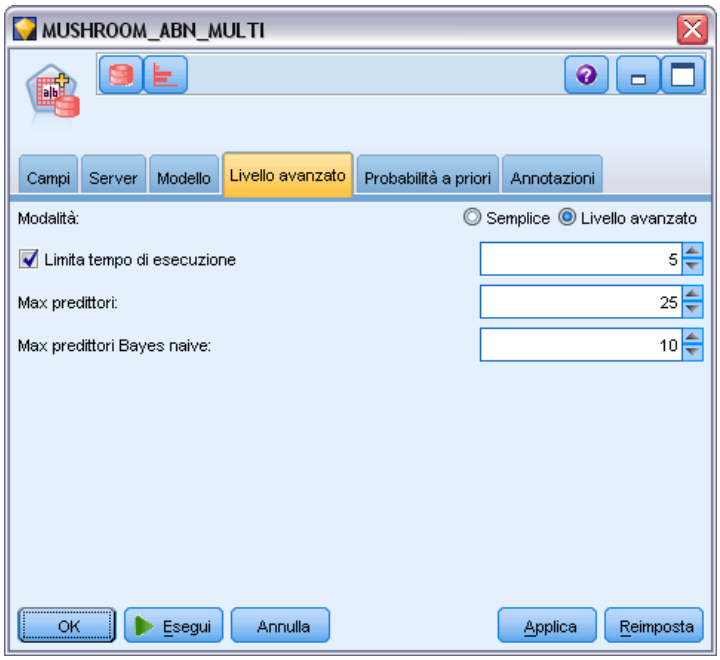
È possibile scegliere fra tre diverse modalità di creazione del modello.

- **Multifunzione.** Crea e confronta diversi modelli, inclusi un modello Bayes naive e modelli di probabilità dei prodotti a funzione singola e multifunzione. Si tratta della modalità più completa e in genere comporta tempi di elaborazione maggiori. Le regole vengono prodotte solo se risulta che il modello a funzione singola è il migliore. Se si sceglie un modello multifunzione o Bayes naive, non vengono prodotte regole.

- **Funzione singola.** Crea un albero decisionale semplificato basato su un insieme di regole. Ogni regola contiene una condizione e le probabilità associate a ciascun risultato. Le regole si escludono a vicenda e vengono fornite in un formato leggibile, offrendo un significativo vantaggio rispetto ai modelli Bayes naive e multifunzione.
- **Bayes naive.** Crea un singolo modello Bayes naive e lo confronta con l'a priori del campione globale (la distribuzione di valori obiettivo nel campione globale). Il modello Bayes naive viene prodotto come output solo se risulta essere un predittore migliore dei valori obiettivo rispetto all'apriori globale. Altrimenti come output non viene prodotto alcun modello.

Opzioni avanzate di Bayes adattivo

Figura 4-8
Opzioni avanzate di Bayes adattivo



Limita tempo di esecuzione. Selezionare questa opzione per specificare un tempo di creazione massimo in minuti. Ciò consente di produrre modelli in tempi più brevi, sebbene ne possa risultare una minore precisione. A ciascun passaggio importante del processo di modellazione, l'algoritmo controlla, prima di continuare, se sarà in grado di completare il passaggio successivo entro l'intervallo di tempo specificato e restituisce il miglior modello disponibile al raggiungimento del limite.

Max predittori. Questa opzione consente di limitare la complessità del modello e migliorare le performance limitando il numero di predittori utilizzati. I predittori vengono classificati in base alla misura MDL della loro correlazione all'obiettivo, come misura della probabilità di essere inclusi nel modello.

Max predittori Bayes naive. Questa opzione specifica il numero massimo di predittori da utilizzare nel modello Bayes naive.

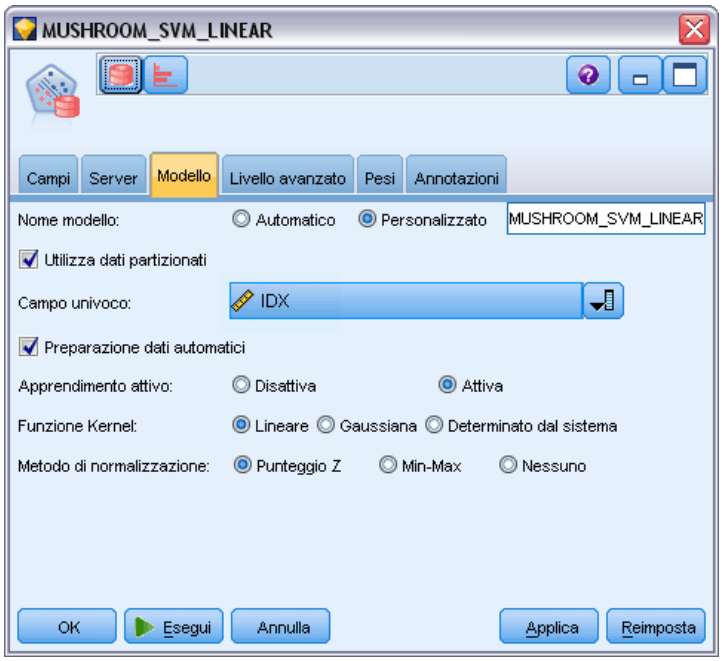
Support Vector Machine Oracle (SVM)

SVM (Support Vector Machine) è un algoritmo di classificazione e regressione che utilizza la teoria di apprendimento automatico per ottimizzare la precisione predittiva senza sovradattare i dati. SVM utilizza una trasformazione non lineare opzionale dei dati di addestramento, seguita dalla ricerca di equazioni di regressione nei dati trasformati per la separazione delle classi (per gli obiettivi categoriali) o l'adattamento dell'obiettivo (per gli obiettivi continui). L'implementazione Oracle di SVM consente di creare i modelli utilizzando uno dei due kernel disponibili, ovvero il kernel lineare o gaussiano. Il kernel lineare omette completamente la trasformazione non lineare, in modo che il modello prodotto risulti essenzialmente un modello di regressione.

Per ulteriori informazioni, consultare i documenti *Oracle Data Mining Application Developer's Guide* e *Oracle Data Mining Concepts*.

Opzioni del modello SVM Oracle

Figura 4-9
Opzioni del modello 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.

Campo univoco. Specifica il campo utilizzato per identificare ciascun caso in modo univoco. Tale campo può coincidere, per esempio, con un campo ID quale *IDCliente*. IBM® SPSS® Modeler impone che questo campo chiave debba essere numerico.

Nota: questo campo è opzionale per tutti i nodi Oracle, ad eccezione dei nodi Bayes adattivi, O-Cluster e Apriori Oracle.

Preparazione dati automatici. (solo 11*g*) Attiva (default) o disattiva la modalità di preparazione dei dati automatici di Oracle Data Mining. Se la casella è selezionata, ODM esegue automaticamente la trasformazione dei dati richiesti dall'algoritmo. Per ulteriori informazioni, vedere *Oracle Data Mining Concepts*.

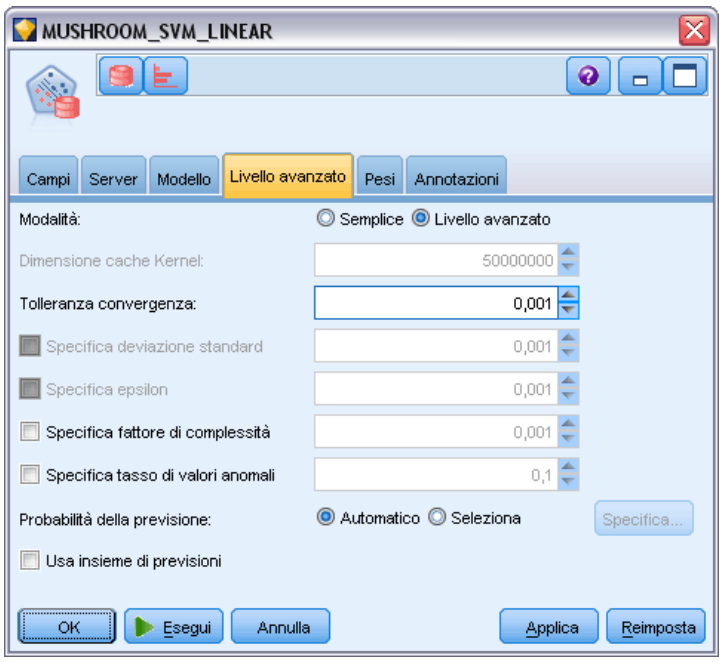
Apprendimento attivo. Fornisce un modo per gestire insiemi di creazione di grandi dimensioni. Mediante l'apprendimento attivo, l'algoritmo crea un modello iniziale basato su un piccolo esempio prima di applicarlo all'insieme di dati di addestramento completo e aggiorna in modo incrementale il campione e il modello in base ai risultati. Il ciclo viene ripetuto finché il modello converge sui dati di addestramento o finché non viene raggiunto il numero massimo di vettori di supporto consentiti.

Funzione Kernel. Selezionare Lineare o Gaussiana, oppure lasciare il valore di default Determinato dal sistema per consentire al sistema di scegliere il kernel più adatto. I kernel gaussiani sono in grado di apprendere relazioni più complesse, ma richiedono in genere tempi di elaborazione maggiori. Può essere opportuno iniziare con il kernel lineare, per poi passare al gaussiano solo se il kernel lineare non riesce a trovare un buon adattamento. Questa situazione si verifica in genere con i modelli di regressione, in cui la scelta del kernel ha un'importanza maggiore. Si noti, inoltre, che i modelli SVM creati con il kernel gaussiano non possono essere visualizzati in SPSS Modeler. I modelli creati con il kernel lineare possono essere visualizzati in SPSS Modeler esattamente come i modelli di regressione standard.

Metodo di normalizzazione. Specifica il metodo di normalizzazione per i campi obiettivo e di input continuo. È possibile scegliere Punteggio Z, Min-Max o Nessuno. Oracle esegue automaticamente la normalizzazione se è selezionata la casella di controllo Preparazione dati automatici. Per impostare manualmente il metodo di normalizzazione, deselezionare la casella di controllo.

Opzioni avanzate di SVM Oracle

Figura 4-10
Opzioni avanzate SVM



Dimensioni cache Kernel. Specifica la dimensione in byte della cache da utilizzare per l'archiviazione dei kernel calcolati durante l'operazione di creazione. Com'è facilmente intuibile, cache di dimensioni maggiori consentono tempi di creazione più rapidi. L'impostazione di default è 50 MB.

Tolleranza convergenza. Specifica il valore di tolleranza consentito prima della terminazione per la creazione del modello. Questo valore deve essere compreso tra 0 e 1. L'impostazione di default è 0.001. Valori maggiori consentono tempi di creazione più rapidi ma producono modelli meno precisi.

Specifica deviazione standard. Specifica il parametro di deviazione standard utilizzato dal kernel gaussiano. Questo parametro incide sul rapporto tra la complessità del modello e la possibilità di essere generalizzato ad altri insiemi di dati (sovradattando e sottoadattando i dati). Valori di deviazione standard maggiori favoriscono il sottoadattamento. Questo parametro viene calcolato di default a partire dai dati di addestramento.

Specifica epsilon. Nei modelli di regressione, specifica il valore dell'intervallo dell'errore consentito nella creazione di modelli senza rilevamento epsilon. In altre parole, distingue errori di piccola portata (che vengono ignorati) da errori più gravi (che non vengono ignorati). Il valore deve essere compreso tra 0 e 1 e viene calcolato di default dai dati di addestramento.

Specifica fattore di complessità. Specifica il fattore di complessità, che bilancia il rapporto tra errore del modello (misurato a fronte dei dati di addestramento) e complessità del modello, per evitare il sovradattamento o il sottoadattamento dei dati. Valori maggiori comportano un livello di penalità più alto per gli errori, con un maggiore rischio di sovradattamento dei dati; valori minori, invece, comportano un livello di penalità più basso e possono portare al sottoadattamento.

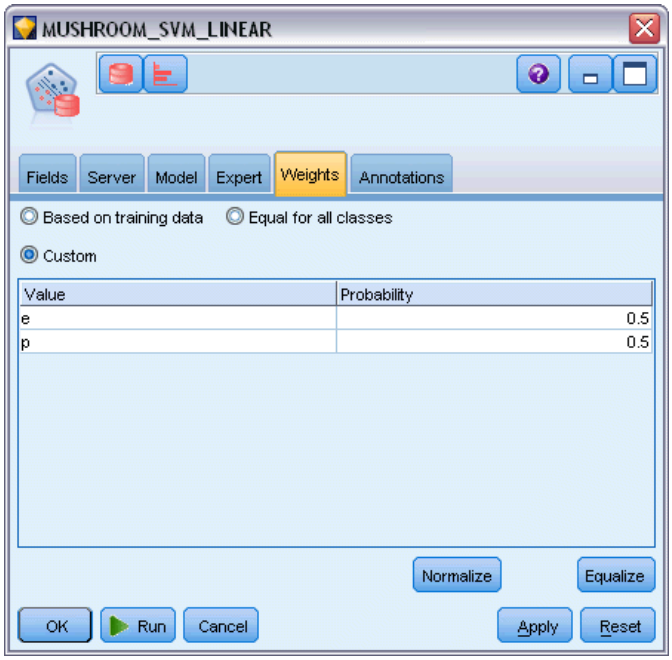
Specifica tasso di valori anomali. Specifica il tasso desiderato di valori anomali nei dati di addestramento. Valido solo per modelli SVM a una classe. Non può essere utilizzata insieme all'impostazione **Specifica fattore di complessità**.

Probabilità della previsione. Consente al modello di includere la probabilità di una previsione corretta di un possibile risultato del campo obiettivo. Per abilitare questa funzione, scegliere Seleziona, fare clic sul pulsante Specifica, scegliere uno dei risultati possibili e fare clic su Inserisci.

Usa insieme di previsioni. Genera una tabella di tutti i risultati possibili relativi a tutti gli esiti possibili del campo obiettivo.

Opzioni Pesi di SVM Oracle

Figura 4-11
Opzioni pesi SVM



In un modello di classificazione, i pesi consentono di specificare l'importanza relativa dei diversi valori di destinazione possibili. Questo può essere utile, per esempio, se i punti dei dati di addestramento non sono distribuiti in modo realistico tra le categorie. I pesi consentono di applicare una distorsione al modello in modo da compensare le categorie meno rappresentate nei dati. L'incremento del peso di un valore di destinazione dovrebbe aumentare la percentuale di previsioni corrette per quella categoria.

Esistono tre metodi per impostare i pesi:

- **In base ai dati di addestramento.** Questa è l'opzione di default. I pesi si basano sulle frequenze relative delle categorie nei dati di addestramento.

- **Uguali per tutte le classi.** I pesi per tutte le categorie sono definiti come 1/*k*, dove *k* è il numero di categorie obiettivo.
- **Personalizzato.** È possibile specificare pesi personalizzati. L'impostazione dei valori iniziali per i pesi è uguale per tutte le classi. È possibile impostare i pesi per le singole categorie su valori definiti dall'utente. Per regolare il peso di una categoria specifica, selezionare la cella Peso nella tabella corrispondente alla categoria desiderata, eliminare il contenuto della cella e immettere il valore desiderato.

La somma devi pesi di tutte le categorie deve essere uguale a 1,0. Se non assommano a 1.0 viene visualizzato un avviso e viene offerta la possibilità di normalizzare automaticamente i valori. Questa modifica automatica mantiene le proporzioni tra le varie categorie e al contempo applica il vincolo di peso. 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.

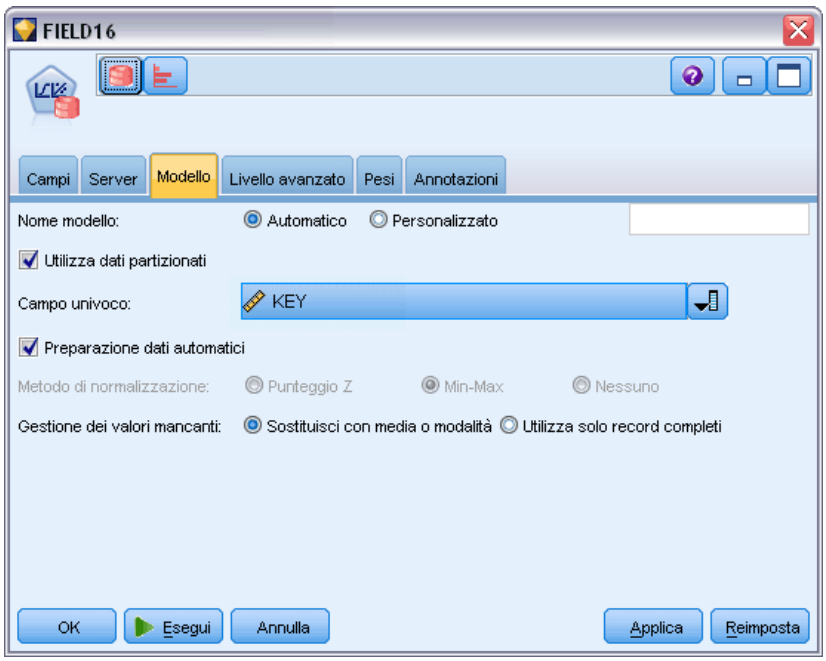
Modelli lineari generalizzati Oracle (GLM)

(11g soltanto) I modelli lineari generalizzati allentano le supposizioni restrittive effettuate dai modelli lineari. Tali supposizioni includono, per esempio, le supposizioni che la variabile obiettivo abbia una distribuzione normale e che l'effetto dei predittori su tale variabile sia lineare per natura. Un modello lineare generalizzato è adatto per le previsioni in cui è probabile che la distribuzione dell'obiettivo sia non normale, per esempio una distribuzione multinomiale o di Poisson. Analogamente, un modello lineare generalizzato è utile nei casi in cui è probabile che la relazione o il collegamento tra i predittori e l'obiettivo siano non-lineari.

Per ulteriori informazioni, consultare i documenti *Oracle Data Mining Application Developer's Guide* e *Oracle Data Mining Concepts*.

Opzioni del modello GLM Oracle

Figura 4-12
Opzioni modello GLM



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.

Campo univoco. Specifica il campo utilizzato per identificare ciascun caso in modo univoco. Tale campo può coincidere, per esempio, con un campo ID quale *IDCliente*. IBM® SPSS® Modeler impone che questo campo chiave debba essere numerico.

Nota: questo campo è opzionale per tutti i nodi Oracle, ad eccezione dei nodi Bayes adattivi, O-Cluster e Apriori Oracle.

Preparazione dati automatici. (solo 11*g*) Attiva (default) o disattiva la modalità di preparazione dei dati automatici di Oracle Data Mining. Se la casella è selezionata, ODM esegue automaticamente la trasformazione dei dati richiesti dall'algoritmo. Per ulteriori informazioni, vedere *Oracle Data Mining Concepts*.

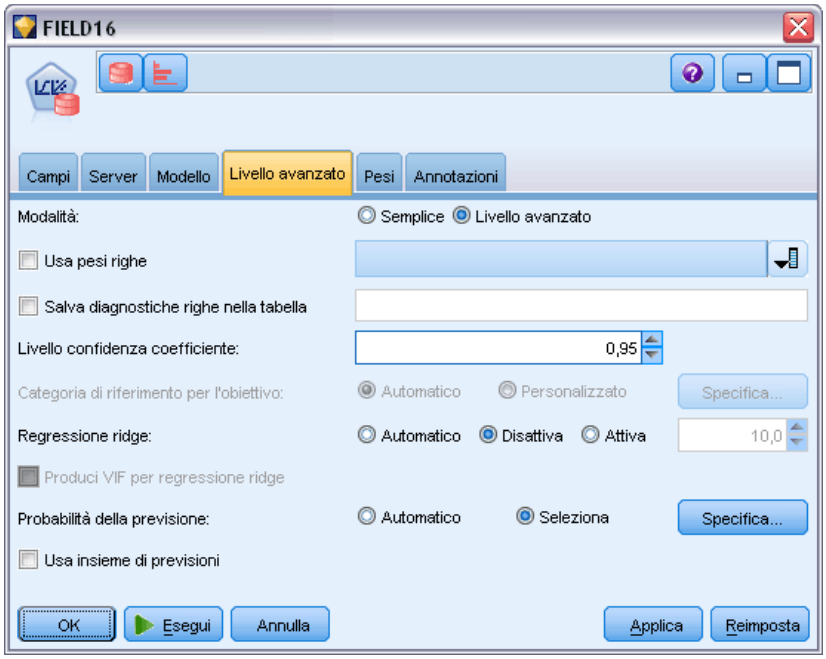
Metodo di normalizzazione. Specifica il metodo di normalizzazione per i campi obiettivo e di input continuo. È possibile scegliere Punteggio Z, Min-Max o Nessuno. Oracle esegue automaticamente la normalizzazione se è selezionata la casella di controllo Preparazione dati automatici. Per impostare manualmente il metodo di normalizzazione, deselezionare la casella di controllo.

Gestione dei valori mancanti. Specifica come elaborare i valori mancanti nei dati di input:

- Sostituisci con media o modalità sostituisce i valori mancanti degli attributi numerici con il valore della media e sostituisce i valori mancanti degli attributi categoriali con la modalità.
- Utilizza solo record completi ignora i record con valori mancanti.

Opzioni avanzate di GLM Oracle

Figura 4-13
Opzioni avanzate GLM



Usa pesi righe. Selezionare questa casella per attivare l'elenco a discesa adiacente da dove è possibile selezionare una colonna contenente un fattore di ponderazione per le righe.

Salva diagnostiche righe nella tabella. Selezionare questa casella di controllo per attivare il campo di testo adiacente in cui è possibile specificare il nome di una tabella contenente diagnostiche a livello di riga.

Livello confidenza coefficiente. Il grado di certezza, da 0,0 a 1,0, che il valore previsto per l'obiettivo rientrerà in un intervallo di confidenza calcolato dal modello. I limiti di confidenza vengono restituiti con le statistiche dei coefficienti.

Categoria di riferimento per l'obiettivo. Selezionare Personalizzato per scegliere un valore del campo obiettivo da utilizzare come categoria di riferimento oppure lasciare il valore di default Auto .

Regressione ridge. La regressione ridge è una tecnica che compensa nel caso in cui il livello di correlazione nelle variabili sia troppo elevato. È possibile utilizzare l'opzione Automatico per consentire all'algoritmo di controllare l'utilizzo di questa tecnica, oppure è possibile controllarlo manualmente mediante le opzioni Disattiva e Attiva. Se si sceglie di attivare manualmente la regressione ridge, è possibile ignorare il valore di default del sistema per il parametro ridge specificando un valore nel campo adiacente.

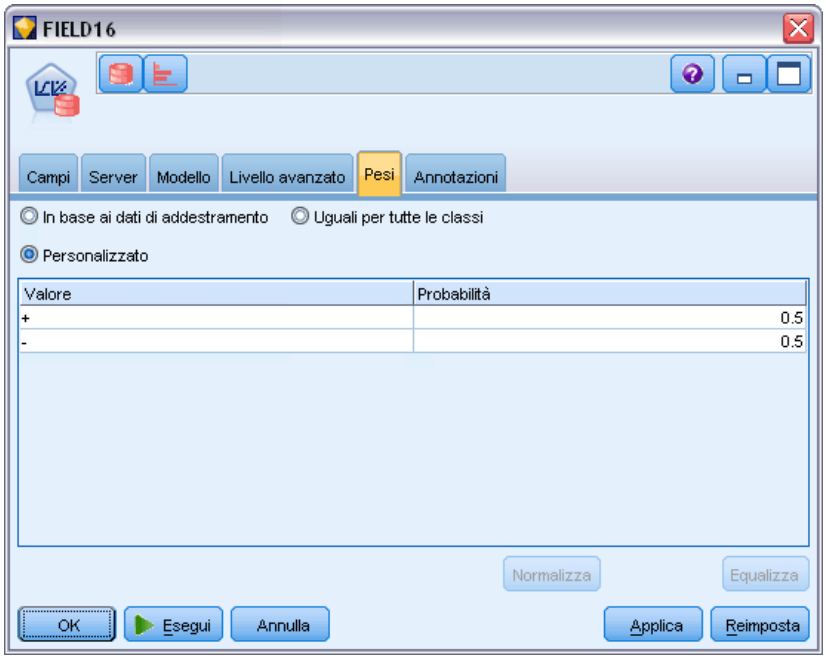
Produci VIF per regressione ridge. Selezionare questa casella se si desidera produrre delle statistiche VIF (Variance Inflation Factor, fattore di inflazione della varianza) quando viene utilizzata la tecnica ridge per la regressione lineare.

Probabilità della previsione. Consente al modello di includere la probabilità di una previsione corretta di un possibile risultato del campo obiettivo. Per abilitare questa funzione, scegliere Seleziona, fare clic sul pulsante Specifica, scegliere uno dei risultati possibili e fare clic su Inserisci.

Usa insieme di previsioni. Genera una tabella di tutti i risultati possibili relativi a tutti gli esiti possibili del campo obiettivo.

Opzioni Pesi di GLM Oracle

Figura 4-14
Opzioni pesi GLM



In un modello di classificazione, i pesi consentono di specificare l'importanza relativa dei diversi valori di destinazione possibili. Questo può essere utile, per esempio, se i punti dei dati di addestramento non sono distribuiti in modo realistico tra le categorie. I pesi consentono di applicare una distorsione al modello in modo da compensare le categorie meno rappresentate nei dati. L'incremento del peso di un valore di destinazione dovrebbe aumentare la percentuale di previsioni corrette per quella categoria.

Esistono tre metodi per impostare i pesi:

- **In base ai dati di addestramento.** Questa è l'opzione di default. I pesi si basano sulle frequenze relative delle categorie nei dati di addestramento.
- **Uguali per tutte le classi.** I pesi per tutte le categorie sono definiti come $1/k$, dove k è il numero di categorie obiettivo.
- **Personalizzato.** È possibile specificare pesi personalizzati. L'impostazione dei valori iniziali per i pesi è uguale per tutte le classi. È possibile impostare i pesi per le singole categorie su valori definiti dall'utente. Per regolare il peso di una categoria specifica, selezionare la cella

Peso nella tabella corrispondente alla categoria desiderata, eliminare il contenuto della cella e immettere il valore desiderato.

La somma devi pesi di tutte le categorie deve essere uguale a 1,0. Se non assommano a 1.0 viene visualizzato un avviso e viene offerta la possibilità di normalizzare automaticamente i valori. Questa modifica automatica mantiene le proporzioni tra le varie categorie e al contempo applica il vincolo di peso. 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.

Albero decisionale Oracle

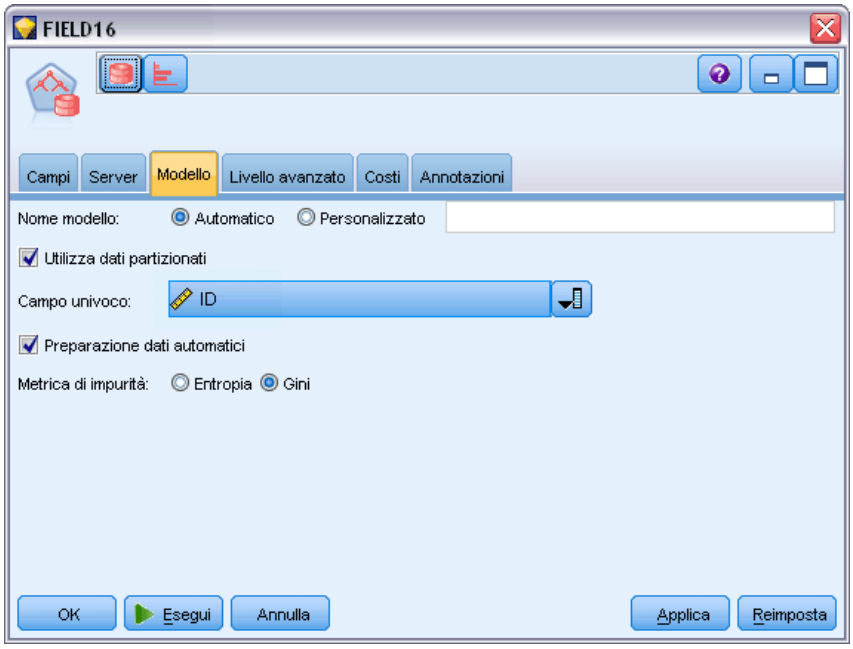
Oracle Data Mining offre una classica funzionalità Albero decisionale, basata sul diffuso algoritmo Alberi di regressione e classificazione. Il modello Albero decisionale ODM contiene informazioni complete sui singoli nodi, fra cui criterio di suddivisione, supporto e confidenza. È possibile visualizzare per intero la Regola per ciascun nodo; inoltre, per ogni nodo viene fornito un attributo surrogato, da utilizzare come sostituto quando si applica il modello a un caso in cui mancano dei valori.

Gli alberi decisionali sono molto diffusi in quanto sono applicabili universalmente oltre a essere facili da utilizzare e capire. Gli alberi decisionali passano in rassegna tutti i potenziali input di attributi alla ricerca del "divisore", vale a dire del punto di taglio degli attributi (per esempio, AGE > 55) che consente di suddividere i record di dati a valle in popolazioni più omogenee. Dopo ogni decisione di divisione, ODM ripete il processo espandendo tutto l'albero verso l'esterno e creando "foglie" terminali che rappresentano gruppi di record, elementi o persone simili. Guardando verso il basso dal nodo radice dell'albero (per esempio la popolazione totale), gli alberi decisionali forniscono regole leggibili di istruzioni IF A, then B. Queste regole dell'albero decisionale forniscono anche il supporto e la confidenza per ciascun nodo dell'albero.

Mentre le reti di Bayes adattivi possono anche fornire regole brevi e semplici, utili a offrire spiegazioni in merito alle singole previsioni, gli alberi decisionali forniscono regole ODM complete per ciascuna decisione di divisione. Gli alberi decisionali sono anche utili per sviluppare profili dettagliati dei clienti migliori, dei pazienti sani, dei fattori associati alla frode, e così via.

Opzioni della scheda Modello per il nodo Albero decisionale

Figura 4-15
Opzioni della scheda Modello di albero decisionale



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.

Campo univoco. Specifica il campo utilizzato per identificare ciascun caso in modo univoco. Tale campo può coincidere, per esempio, con un campo ID quale *IDCliente*. IBM® SPSS® Modeler impone che questo campo chiave debba essere numerico.

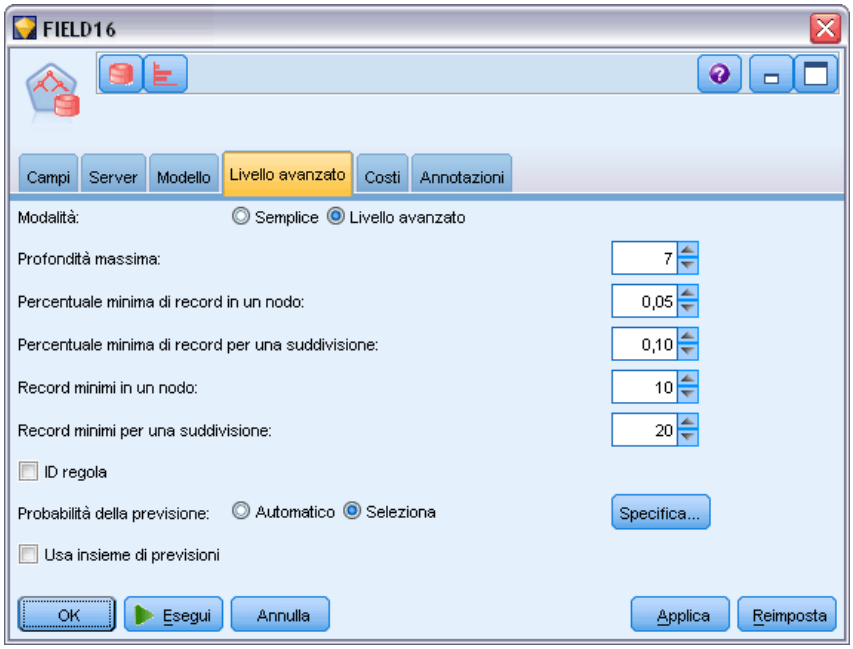
Nota: questo campo è opzionale per tutti i nodi Oracle, ad eccezione dei nodi Bayes adattivi, O-Cluster e Apriori Oracle.

Preparazione dati automatici. (solo 11*g*) Attiva (default) o disattiva la modalità di preparazione dei dati automatici di Oracle Data Mining. Se la casella è selezionata, ODM esegue automaticamente la trasformazione dei dati richiesti dall'algoritmo. Per ulteriori informazioni, vedere *Oracle Data Mining Concepts*.

Metrica di impurità. Specifica la metrica utilizzata per cercare la migliore domanda di test per la suddivisione dei dati nei singoli nodi. Il divisore e il valore di suddivisione migliori sono quelli che risultano nel maggior incremento dell'omogeneità del valore di destinazione per le entità presenti nel nodo. L'omogeneità viene misurata in base a un tipo di metrica. Sono supportate le metriche **gini** ed **entropia**.

Opzioni avanzate Albero decisionale

Figura 4-16
Opzioni avanzate Albero decisionale



Profondità massima. Imposta la profondità massima del modello di albero da creare.

Percentuale minima di record in un nodo. Imposta la percentuale del numero minimo di record per nodo.

Percentuale minima di record per una suddivisione. Imposta il numero minimo di record in un nodo genitore, espresso come valore percentuale del numero totale dei record utilizzati per l'addestramento del modello. Se il numero dei record è inferiore a questo valore percentuale, il sistema non esegue alcuna suddivisione.

Record minimi in un nodo. Imposta il numero minimo di record restituiti.

Record minimi per una suddivisione. Imposta il numero minimo di record in un nodo genitore, espresso sotto forma di valore. Se il numero dei record è inferiore a questo valore, il sistema non esegue alcuna suddivisione.

ID regola. Se selezionata, include nel modello una stringa per identificare il nodo dell'albero in cui viene eseguita una determinata suddivisione.

Probabilità della previsione. Consente al modello di includere la probabilità di una previsione corretta di un possibile risultato del campo obiettivo. Per abilitare questa funzione, scegliere Seleziona, fare clic sul pulsante Specifica, scegliere uno dei risultati possibili e fare clic su Inserisci.

Usa insieme di previsioni. Genera una tabella di tutti i risultati possibili relativi a tutti gli esiti possibili del campo obiettivo.

O-Cluster Oracle

L'algoritmo O-Cluster Oracle consente di identificare i raggruppamenti che si creano spontaneamente all'interno di una popolazione di dati. Il raggruppamento cluster a partizioni ortogonali, O-Cluster, è un algoritmo di raggruppamento proprietario di Oracle che consente di creare un modello di raggruppamento gerarchico basato su una griglia, vale a dire, crea partizioni ortogonali (parallele all'asse) nello spazio dell'attributo di input. Questo algoritmo funziona in modo ricorsivo. La struttura gerarchica risultante rappresenta una griglia irregolare che suddivide lo spazio dell'attributo in raggruppamenti a tasselli.

L'algoritmo O-Cluster gestisce sia gli attributi numerici sia gli attributi categoriali e ODM seleziona automaticamente le definizioni di raggruppamenti cluster migliori. ODM fornisce informazioni dettagliate sui cluster, le relative regole e i valori del baricentro e può esser utilizzato per calcolare il punteggio di una popolazione in base all'appartenenza al cluster.

Opzioni del modello O-Cluster

Figura 4-17
opzioni del modello O-Cluster

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.

Campo univoco. Specifica il campo utilizzato per identificare ciascun caso in modo univoco. Tale campo può coincidere, per esempio, con un campo ID quale *IDCliente*. IBM® SPSS® Modeler impone che questo campo chiave debba essere numerico.

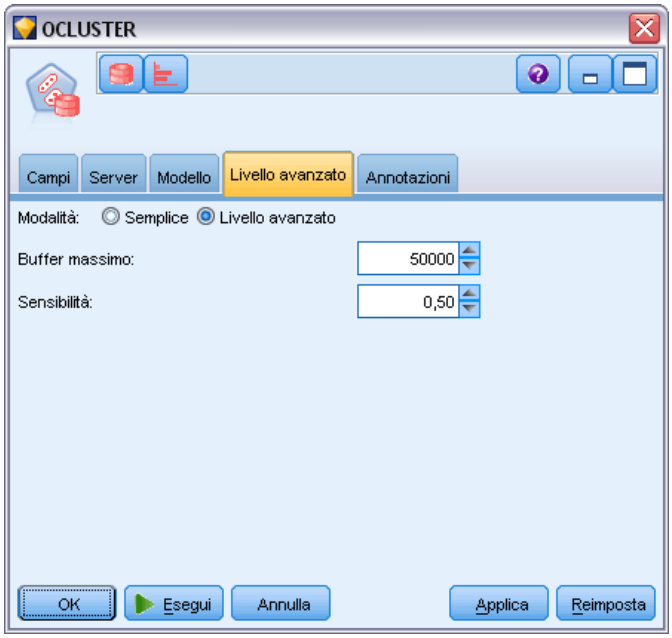
Nota: questo campo è opzionale per tutti i nodi Oracle, ad eccezione dei nodi Bayes adattivi, O-Cluster e Apriori Oracle.

Preparazione dati automatici. (solo 11*g*) Attiva (default) o disattiva la modalità di preparazione dei dati automatici di Oracle Data Mining. Se la casella è selezionata, ODM esegue automaticamente la trasformazione dei dati richiesti dall'algoritmo. Per ulteriori informazioni, vedere *Oracle Data Mining Concepts*.

Numero massimo di cluster. Imposta il numero massimo di cluster generati.

Opzioni avanzate di O-Cluster

Figura 4-18
opzioni avanzate di O-Cluster



Buffer massimo. Imposta le dimensioni massime del buffer.

Sensibilità. Imposta una frazione che specifica la densità di picco necessaria per la separazione di un nuovo cluster. Il valore frazionale è relativo alla densità uniforme globale.

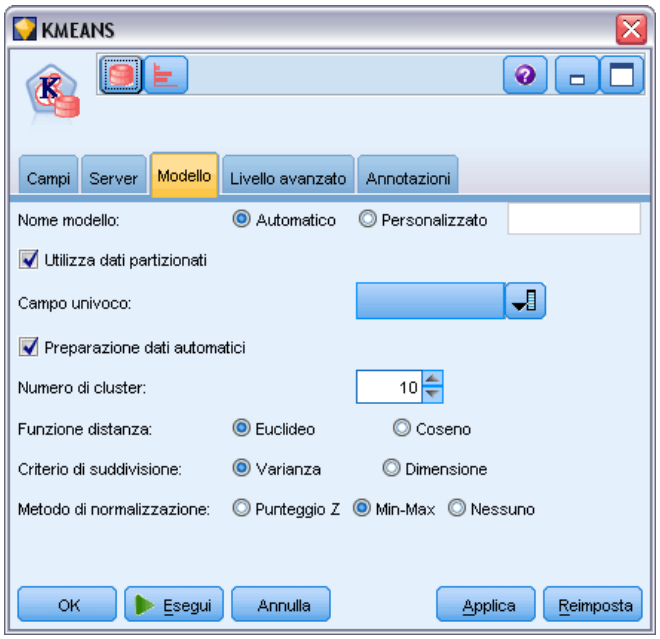
K-Means Oracle

L'algoritmo K-Means Oracle consente di identificare i raggruppamenti che si creano spontaneamente all'interno di una popolazione di dati. L'algoritmo K-Means è un algoritmo di raggruppamento cluster basato sulla distanza che suddivide i dati sotto forma di partizioni in un numero predeterminato di cluster (a condizione che ci sia un numero sufficiente di casi distinti). Gli algoritmi basati sulla distanza fanno affidamento su una metrica di distanza (funzione) per misurare la similarità tra i punti dei dati. I punti dei dati vengono assegnati al cluster più vicino in base alla metrica di distanza utilizzata. ODM fornisce una versione migliorata di K-Means.

L'algoritmo K-Means supporta i cluster gerarchici, gestisce gli attributi numerici e categoriali e suddivide la popolazione nel numero di cluster specificati dall'utente. ODM fornisce informazioni dettagliate sui cluster, le relative regole e i valori del baricentro e può esser utilizzato per calcolare il punteggio di una popolazione in base all'appartenenza al cluster.

Opzioni del modello K-Means

Figura 4-19
opzioni del modello K-Means



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.

Campo univoco. Specifica il campo utilizzato per identificare ciascun caso in modo univoco. Tale campo può coincidere, per esempio, con un campo ID quale *IDCliente*. IBM® SPSS® Modeler impone che questo campo chiave debba essere numerico.

Nota: questo campo è opzionale per tutti i nodi Oracle, ad eccezione dei nodi Bayes adattivi, O-Cluster e Apriori Oracle.

Preparazione dati automatici. (solo 11*g*) Attiva (default) o disattiva la modalità di preparazione dei dati automatici di Oracle Data Mining. Se la casella è selezionata, ODM esegue automaticamente la trasformazione dei dati richiesti dall'algoritmo. Per ulteriori informazioni, vedere *Oracle Data Mining Concepts*.

Numero di cluster. Imposta il numero di cluster generati.

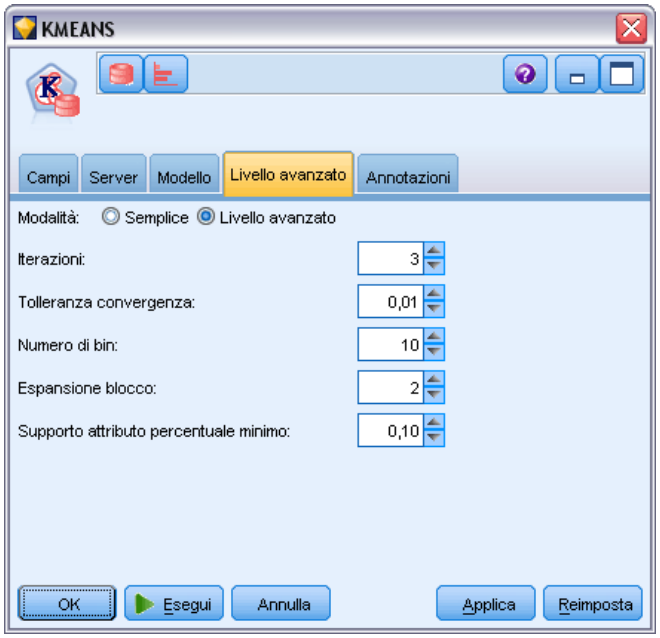
Funzione distanza. Specifica la funzione distanza utilizzata per il raggruppamento cluster K-Means.

Criterio di suddivisione. Specifica il criterio di suddivisione utilizzato per il raggruppamento cluster K-Means.

Metodo di normalizzazione. Specifica il metodo di normalizzazione per i campi obiettivo e di input continuo. È possibile scegliere Punteggio Z, Min-Max o Nessuno.

Opzioni avanzate del nodo K-Means

Figura 4-20
Opzioni avanzate del nodo K-Means



Iterazioni. Specifica il numero di iterazioni per l'algoritmo K-Means.

Tolleranza convergenza. Imposta la tolleranza di convergenza per l'algoritmo K-Means.

Numero di bin. Specifica il numero di bin nell'istogramma dell'attributo prodotto da K-Means. I limiti del bin per i singoli attributi vengono calcolati globalmente sull'intero insieme di dati di addestramento. Il metodo di discretizzazione è ad ampiezza equivalente. Tutti gli attributi hanno lo stesso numero di bin ad eccezione degli attributi con un unico valore, che invece hanno un solo bin.

Espansione blocco. Imposta il fattore di espansione relativo alla memoria allocata per la memorizzazione dei dati dei cluster.

Supporto attributo percentuale minimo. Imposta la frazione dei valori dell'attributo che devono essere non nulli per far sì che l'attributo venga incluso nella descrizione della regola per il cluster. L'impostazione di un valore troppo alto per questo parametro nel caso di dati con valori mancanti può determinare la creazione di regole molto brevi o addirittura vuote.

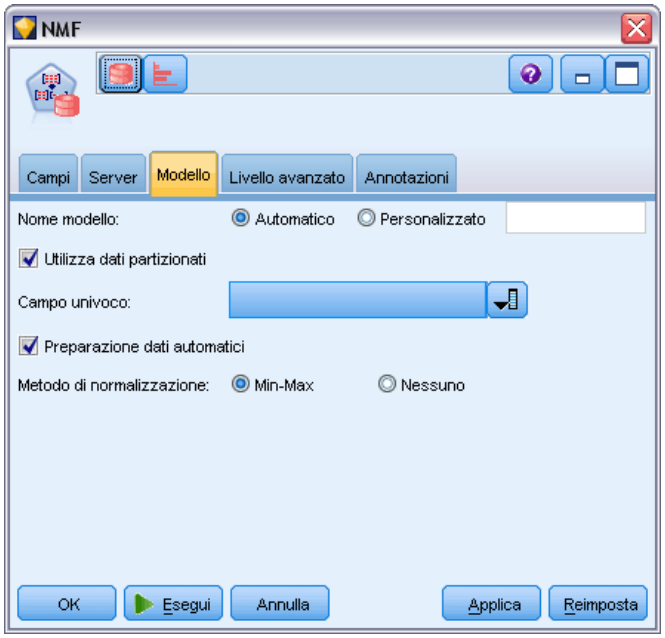
NMF di Oracle (fattorizzazione a matrice non negativa)

L'algoritmo NMF è utile per ridurre un insieme di dati molto grosso in attributi rappresentativi. Concettualmente analogo all'algoritmo di Analisi dei componenti principali (PCA) ma in grado di gestire quantità di attributi maggiori in un modello di rappresentazione additivo, l'algoritmo NMF è un algoritmo di data mining potente e all'avanguardia, che può essere utilizzato in svariati casi.

L'algoritmo NMF può essere utilizzato per ridurre grandi quantità di dati (per esempio dati di testo) in rappresentazioni più piccole e sparse, che riducono la dimensionalità dei dati (le stesse informazioni possono essere conservate utilizzando un numero di variabili molto inferiore). L'output dei modelli NMF può essere analizzato mediante tecniche di apprendimento supervisionato, come le tecniche SVM, o non supervisionato, come le tecniche di raggruppamento tramite cluster. Oracle Data Mining utilizza gli algoritmi NMF e SVM per eseguire il mining di dati di testo non strutturati.

Opzioni del modello NMF

Figura 4-21
opzioni del modello NMF



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.

Campo univoco. Specifica il campo utilizzato per identificare ciascun caso in modo univoco. Tale campo può coincidere, per esempio, con un campo ID quale *IDCliente*. IBM® SPSS® Modeler impone che questo campo chiave debba essere numerico.

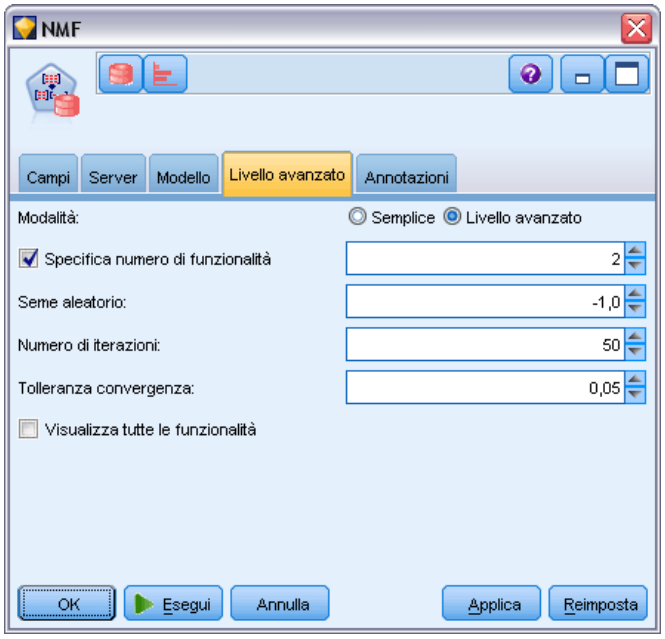
Nota: questo campo è opzionale per tutti i nodi Oracle, ad eccezione dei nodi Bayes adattivi, O-Cluster e Apriori Oracle.

Preparazione dati automatici. (solo 11*g*) Attiva (default) o disattiva la modalità di preparazione dei dati automatici di Oracle Data Mining. Se la casella è selezionata, ODM esegue automaticamente la trasformazione dei dati richiesti dall'algoritmo. Per ulteriori informazioni, vedere *Oracle Data Mining Concepts*.

Metodo di normalizzazione. Specifica il metodo di normalizzazione per i campi obiettivo e di input continuo. È possibile scegliere Punteggio Z, Min-Max o Nessuno. Oracle esegue automaticamente la normalizzazione se è selezionata la casella di controllo Preparazione dati automatici. Per impostare manualmente il metodo di normalizzazione, deselezionare la casella di controllo.

Opzioni avanzate NMF

Figura 4-22
opzioni avanzate NMF



Specifica numero di funzionalità. Specifica il numero delle funzionalità da estrarre.

Seme aleatorio. Imposta il seme aleatorio per l'algoritmo NMF.

Numero di iterazioni. Imposta il numero delle iterazioni per l'algoritmo NMF.

Tolleranza convergenza. Imposta la tolleranza di convergenza per l'algoritmo NMF.

Visualizza tutte le funzionalità. Consente di visualizzare ID e confidenza per tutte le funzionalità e non solo per la funzionalità migliore.

Apriori Oracle

L'algoritmo Apriori scopre le regole di associazione presenti nei dati. Ad esempio, "se un cliente acquista un rasoio e una lozione dopobarba, esiste una confidenza dell'80% che comprerà poi la schiuma da barba". Il problema di mining di associazione può essere scomposto in due sottoproblemi:

- Trovare tutte le combinazioni di elementi, denominate insiemi di elementi frequenti, il cui supporto sia superiore al valore di supporto minimo.
- Utilizzare gli insiemi di elementi frequenti per generare le regole desiderate. Il concetto è il seguente: se, per esempio, ABC e BC sono frequenti, la regola "A implica BC" è vera se il rapporto tra support(ABC) e support(BC) è grande almeno tanto quanto la confidenza minima. Si noti che la regola avrà un supporto minimo in quanto ABCD è frequente. L'associazione ODM supporta solo le regole conseguenti singole (ABC implica D).

Il numero degli insiemi di elementi frequenti è determinato dai parametri di supporto minimo. Il numero delle regole generate è determinato dal numero degli insiemi di elementi frequenti e dal parametro di confidenza. Se il parametro di confidenza è stato impostato su un valore troppo alto, è possibile che nel modello di associazione siano presenti degli insiemi di elementi frequenti ma nessuna regola.

ODM utilizza un'implementazione basata su SQL dell'algoritmo Apriori. La generazione di candidati e i passaggi di conteggio del supporto vengono implementati mediante query SQL. Non vengono invece utilizzate le strutture di dati in memoria specializzate. Le query SQL vengono perfezionate in modo da essere eseguite efficacemente nel server del database, utilizzando diversi suggerimenti.

Opzioni dei campi Apriori

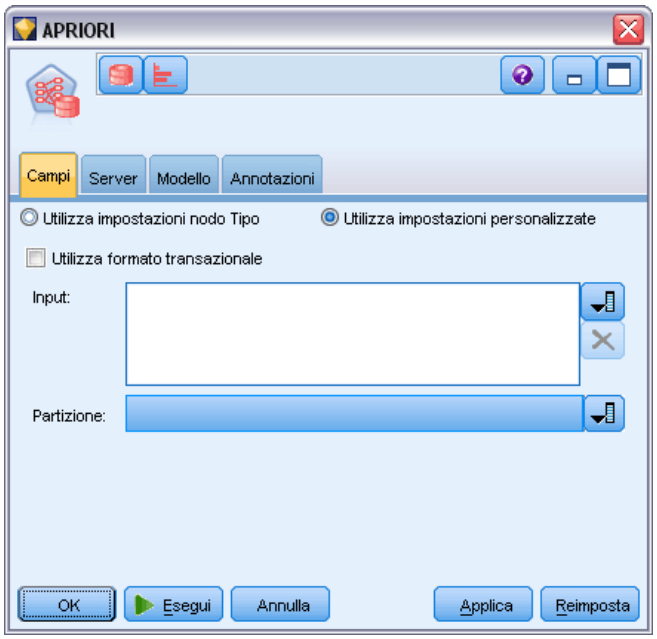
In tutti i nodi Modelli è disponibile una scheda Campi nella quale è possibile specificare i campi da utilizzare per la creazione del modello.

Per poter generare un modello Apriori, è necessario prima specificare i campi da utilizzare come elementi rilevanti nella creazione di modelli di associazione.

Utilizza impostazioni nodo Tipo. Questa opzione indica al nodo di utilizzare le informazioni sui campi da un nodo Tipo a monte. È l'impostazione 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 rimanenti nella finestra di dialogo, che dipenderanno dall'utilizzo o meno del formato transazionale.

Figura 4-23
Impostazioni personalizzate di default dei campi



Se *non si utilizza* il formato transazionale, specificare:

- **Input.** Selezionare i campi 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 validazione della creazione del modello. Per ulteriori informazioni, vedere l'argomento Nodo Partizione in il capitolo 4 in *IBM SPSS Modeler 15 Nodi di input, elaborazione e output*.

Se *si utilizza* il formato transazionale, specificare:

Utilizza formato transazionale. Utilizzare questa opzione se si desidera trasformare i dati da una riga per voce a una riga per caso.

La selezione di questa opzione modifica i controlli dei campi nella parte inferiore di questa finestra di dialogo:

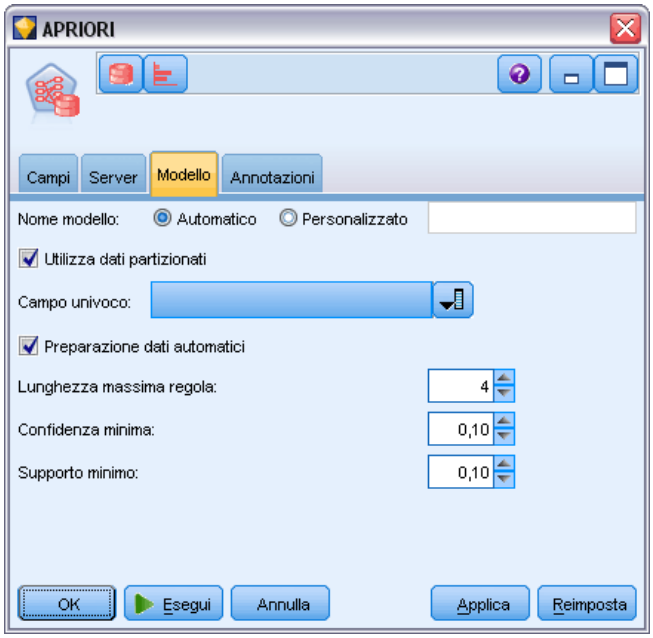
Figura 4-24
Impostazioni dei campi per il formato transazionale

Per il formato transazionale, specificare:

- **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).
- **Contenuto.** Specificare il campo contenuto per il modello. Questo campo contiene l'elemento rilevante nella creazione di modelli di associazione.
- **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 validazione della creazione del modello. Utilizzando un campione per creare 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 di partizione utilizzando nodi Tipo o Partizione, nella scheda Campi di ogni nodo Modelli che utilizza il partizionamento è necessario selezionare un singolo campo di partizione. Se è presente un'unica partizione, verrà utilizzata automaticamente quando si attiva il partizionamento. Per ulteriori informazioni, vedere l'argomento Nodo Partizione in il capitolo 4 in *IBM SPSS Modeler 15 Nodi di input, elaborazione e output*. 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 Apriori

Figura 4-25
opzioni del modello 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.

Campo univoco. Specifica il campo utilizzato per identificare ciascun caso in modo univoco. Tale campo può coincidere, per esempio, con un campo ID quale *IDCliente*. IBM® SPSS® Modeler impone che questo campo chiave debba essere numerico.

Nota: questo campo è opzionale per tutti i nodi Oracle, ad eccezione dei nodi Bayes adattivi, O-Cluster e Apriori Oracle.

Preparazione dati automatici. (solo 11*g*) Attiva (default) o disattiva la modalità di preparazione dei dati automatici di Oracle Data Mining. Se la casella è selezionata, ODM esegue automaticamente la trasformazione dei dati richiesti dall'algoritmo. Per ulteriori informazioni, vedere *Oracle Data Mining Concepts*.

Lunghezza massima regola. Imposta il numero massimo di precondizioni per qualsiasi regola, un intero da 2 a 20. È un modo per limitare la complessità delle regole. Se le regole sono troppo complesse o troppo specifiche, oppure se l'addestramento della propria regola sta richiedendo troppo tempo, provare a diminuire questo valore.

Confidenza minima. Imposta il livello di confidenza minimo, un valore tra 0 e 1. Le regole con una confidenza inferiore rispetto al criterio specificato vengono scartate.

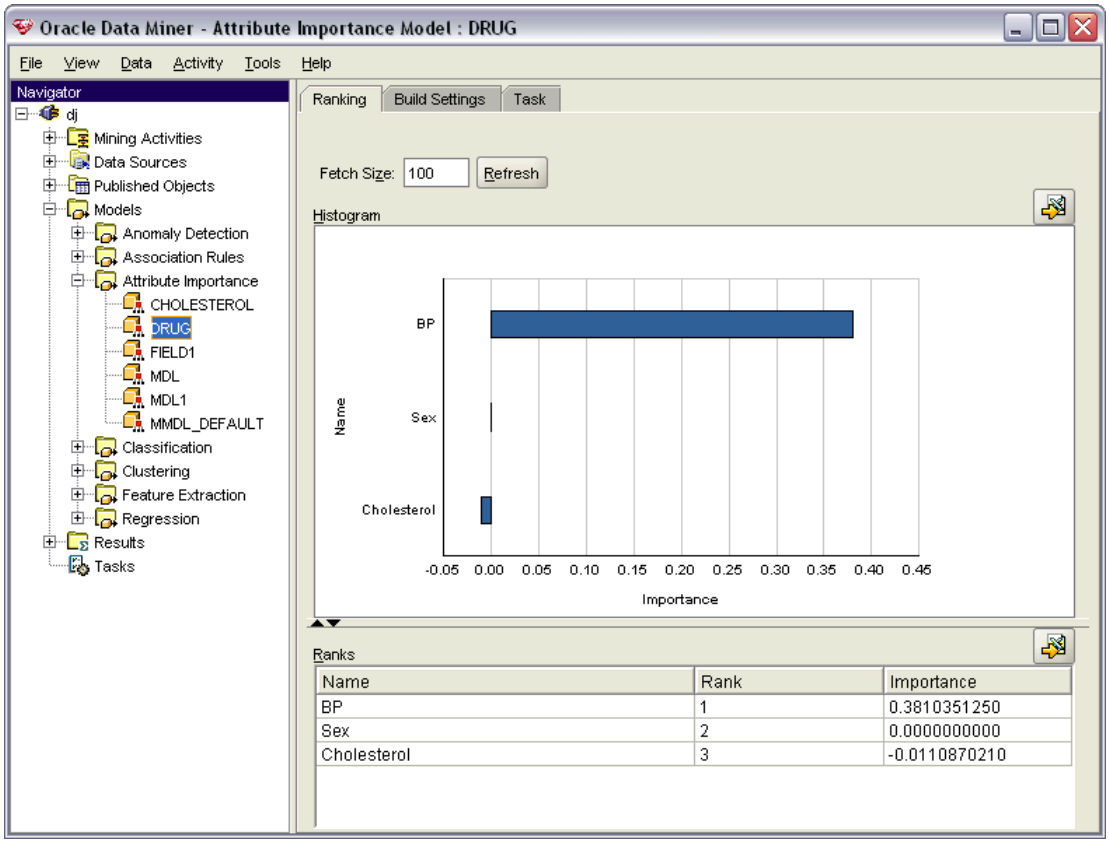
Supporto minimo. Imposta la soglia di supporto minimo, un valore tra 0 e 1. Apriori scopre schemi con una frequenza superiore alla soglia di supporto minimo.

Oracle MDL (Lunghezza descrizione minima)

L'algoritmo di Oracle MDL (Lunghezza descrizione minima) facilita l'identificazione degli attributi che hanno il maggiore impatto su un attributo obiettivo. Spesso, sapere quali sono gli attributi più influenti aiuta a capire e gestire meglio la propria attività e consente di semplificare le attività di creazione dei modelli. Inoltre, questi attributi possono indicare i tipi di dati che è possibile aggiungere per espandere i modelli. L'algoritmo MDL può essere utilizzato, per esempio, per identificare gli attributi del processo più rilevanti per prevedere la qualità di un componente prodotto, i fattori associati al tasso di abbandono o i geni che potrebbero essere coinvolti nella cura di una malattia specifica.

Oracle MDL scarta i campi di input che considera privi di importanza nella previsione dell'obiettivo. Con i campi di input restanti crea quindi un insieme di modelli grezzo associato a un modello Oracle, visibile in Oracle Data Miner. Quando si consulta il modello in Oracle Data Miner, viene visualizzato un grafico che mostra i campi di input rimanenti, in ordine di significatività ai fini della previsione dell'obiettivo.

Figura 4-26
Utilizzo del grafico di Oracle MDL che mostra l'importanza relativa dei campi di input nella previsione di un obiettivo



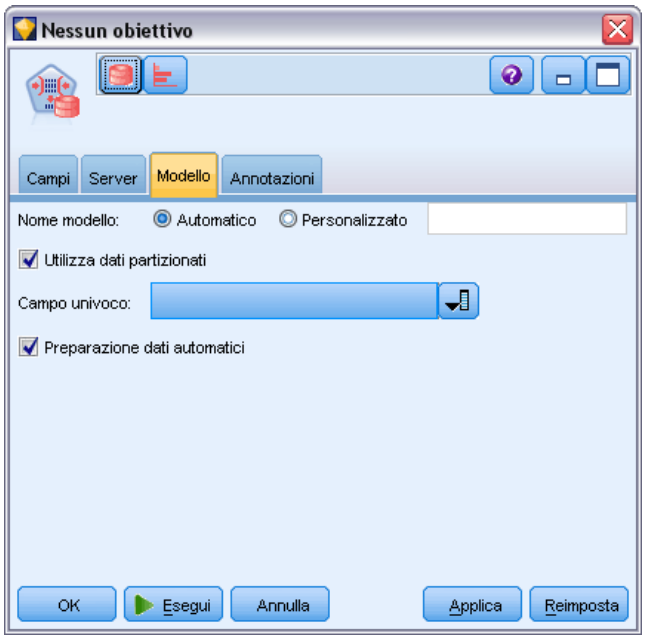
Una classificazione negativa indica rumore. I campi di input classificati zero o meno di zero non contribuiscono alla previsione ed è consigliabile eliminarli dai dati.

Per visualizzare il grafico

- Con il pulsante destro del mouse, fare clic sull'insieme di modelli grezzo nella palette Modelli e scegliere Visualizza.
- Dalla finestra del modello, fare clic sul pulsante per lanciare Oracle Data Miner.
- Connettersi a Oracle Data Miner. Per ulteriori informazioni, vedere l'argomento Oracle Data Miner a pag. 96.
- Nel pannello di navigazione di Oracle Data Miner, espandere Modelli e quindi Importanza attributo.
- Selezionare il modello Oracle desiderato (che avrà lo stesso nome del campo obiettivo specificato in IBM® SPSS® Modeler). Se non si conosce il modello corretto, selezionare la cartella Importanza attributo e cercare un modello in base alla data di creazione.

Opzioni del modello MDL

Figura 4-27
opzioni del modello MDL



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.

Campo univoco. Specifica il campo utilizzato per identificare ciascun caso in modo univoco. Tale campo può coincidere, per esempio, con un campo ID quale *IDCliente*. IBM® SPSS® Modeler impone che questo campo chiave debba essere numerico.

Nota: questo campo è opzionale per tutti i nodi Oracle, ad eccezione dei nodi Bayes adattivi, O-Cluster e Apriori Oracle.

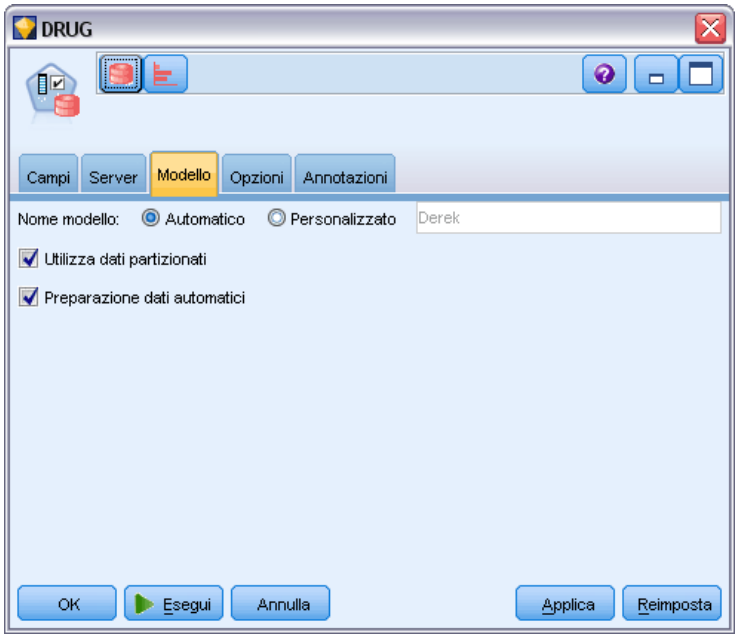
Preparazione dati automatici. (solo 11*g*) Attiva (default) o disattiva la modalità di preparazione dei dati automatici di Oracle Data Mining. Se la casella è selezionata, ODM esegue automaticamente la trasformazione dei dati richiesti dall'algoritmo. Per ulteriori informazioni, vedere *Oracle Data Mining Concepts*.

Importanza attributo Oracle (AI)

Scopo dell'importanza attributo è individuare gli attributi dell'insieme di dati correlati al risultato, e in che misura influiscono sul risultato finale. Il nodo Importanza attributo Oracle analizza dati, individua schemi e prevede risultati con un livello di confidenza associato.

Opzioni modello AI

Figura 4-28
Opzioni modello AI



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 di partizione, questa opzione garantisce che per la creazione del modello verranno utilizzati solo i dati della partizione di addestramento. Per ulteriori informazioni, vedere l'argomento Nodo Partizione in il capitolo 4 in *IBM SPSS Modeler 15 Nodi di input, elaborazione e output*.

Preparazione dati automatici. (solo 11*g*) Attiva (default) o disattiva la modalità di preparazione dei dati automatici di Oracle Data Mining. Se la casella è selezionata, ODM esegue automaticamente la trasformazione dei dati richiesti dall'algoritmo. Per ulteriori informazioni, vedere *Oracle Data Mining Concepts*.

Opzioni di selezione AI

La scheda Opzioni consente di specificare le impostazioni di default per la selezione o l'esclusione dei campi di input nell'insieme di modelli. In seguito è possibile aggiungere il modello a uno stream 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 l'insieme di modelli senza ulteriori modifiche, il che può rivelarsi particolarmente utile ai fini dello script.

Figura 4-29
Opzioni di selezione AI

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.

Scheda Modello dell'insieme di modelli AI

La scheda Modello di un insieme di modelli Oracle AI visualizza il rango e l'importanza di tutti gli input e consente di selezionare i campi da filtrare utilizzando le caselle di controllo nella colonna di sinistra. Quando si esegue lo stream, vengono mantenuti solo i campi selezionati insieme alla previsione dell'obiettivo. Gli altri campi di input vengono scartati. Le selezioni

di default sono basate sulle opzioni specificate nel nodo Modelli, ma è possibile selezionare o deselezionare campi aggiuntivi, se necessario.

Figura 4-30
Insieme di modelli AI

- 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, selezionare l'elemento desiderato dall'elenco accanto al pulsante 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 deselezionare tutti i campi e per accedere alla finestra di dialogo Seleziona campi, che consente di selezionare i campi per rango o per importanza. Per estendere la selezione è possibile anche premere il tasto Maiusc o Ctrl mentre si fa clic sui campi. Per ulteriori informazioni, vedere l'argomento Selezione dei campi per importanza in il capitolo 4 in *IBM SPSS Modeler 15 Nodi Modelli*.
- 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.

Gestione dei modelli Oracle

I modelli Oracle vengono aggiunti alla palette Modelli esattamente come gli altri modelli IBM® SPSS® Modeler e possono essere utilizzati in modo sostanzialmente simile. Tuttavia, esistono alcune importanti differenze, dato che ogni modello Oracle creato in SPSS Modeler fa riferimento a un modello archiviato in un server di database.

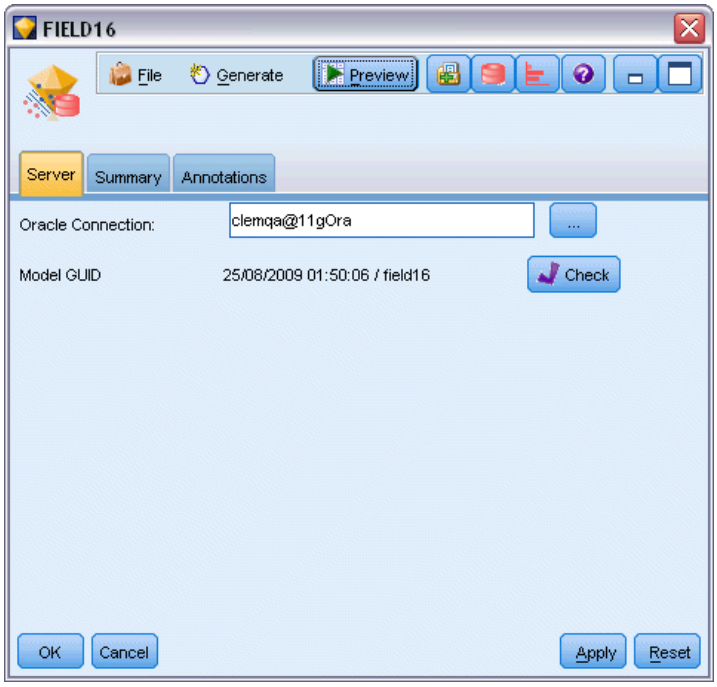
Scheda Server dell'insieme di modelli Oracle

Se si crea un modello ODM tramite IBM® SPSS® Modeler, viene creato un modello in SPSS Modeler e viene creato o sostituito un modello nel database Oracle. Un modello di SPSS Modeler di questo tipo fa riferimento al contenuto di un modello di database archiviato in un server di database. SPSS Modeler consente di eseguire un controllo dell'uniformità archiviando una stringa identica con la **chiave del modello** generato sia nel modello di SPSS Modeler che nel modello Oracle.

La stringa della chiave per ciascun modello Oracle viene visualizzata nella colonna *Informazioni sul modello* della finestra di dialogo Elenca modelli. La stringa della chiave di un modello di SPSS Modeler viene invece visualizzata come Chiave di modello nella scheda Server di un modello SPSS Modeler (se all'interno di uno stream).

È possibile utilizzare il pulsante Controllo della scheda Server di un insieme di modelli per controllare che le chiavi di modello nei modelli SPSS Modeler e Oracle corrispondano. Se in Oracle non è reperibile alcun modello con lo stesso nome o se le chiavi del modello non corrispondono, il modello Oracle è stato eliminato o ricreato dopo la creazione del modello di SPSS Modeler.

Figura 4-31
Opzioni della scheda Server dell'insieme di modelli Oracle



Scheda Riepilogo dell'insieme di modelli Oracle

La scheda Riepilogo di un insieme di modelli 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 di espansione 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, utilizzare il controllo di espansione di un elemento se si desidera nascondere solo i risultati di tale elemento oppure fare clic sul pulsante Comprimi tutto se si desidera nascondere tutti i risultati.

Analisi. Visualizza informazioni sul modello specifico. Se è stato eseguito un nodo Analisi collegato a questo insieme di modelli, in questa sezione verranno visualizzate anche le informazioni sui risultati di tale analisi. Per ulteriori informazioni, vedere l'argomento nodo Analisi in il capitolo 6 in *IBM SPSS Modeler 15 Nodi di input, elaborazione e output*.

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, lo stream utilizzato per creare tale modello, l'utente che lo ha creato, la data di creazione e il tempo trascorso per la creazione del modello.

Scheda Impostazioni dell'insieme di modelli Oracle

La scheda Impostazioni sull'insieme di modelli consente di ignorare l'impostazione di certe opzioni sul nodo Modelli a scopi di calcolo del punteggio.

Albero decisionale Oracle

Utilizza costi di errata classificazione. Determina se utilizzare i costi di errata classificazione nel modello Albero decisionale Oracle. Per ulteriori informazioni, vedere l'argomento Costi classificazione errata a pag. 61.

ID regola. Se selezionata, aggiunge una colonna ID regola al modello Albero decisionale Oracle. L'ID regola identifica il nodo nell'albero in cui viene eseguita una determinata suddivisione.

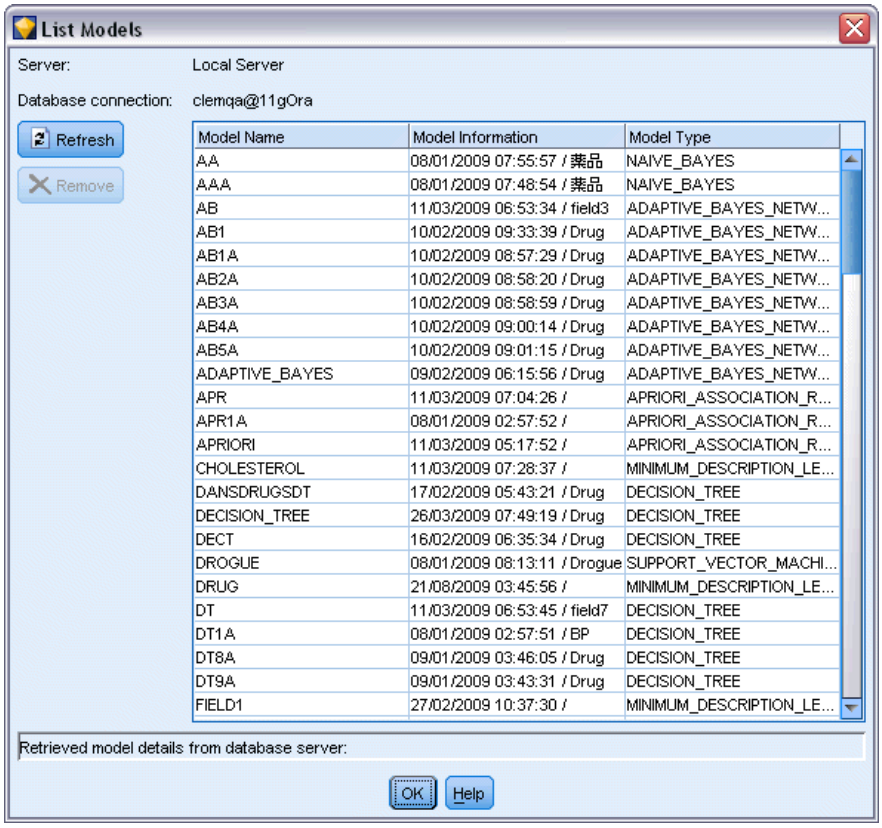
NMF Oracle

Visualizza tutte le funzionalità. Se selezionata, consente di visualizzare ID e confidenza per tutte le funzionalità e non solo per la funzionalità migliore, nel modello NMF Oracle.

Elenco dei modelli Oracle

Il pulsante Elenca modelli Oracle Data Mining avvia una finestra di dialogo in cui sono elencati i modelli di database esistenti e da cui è possibile rimuovere i modelli. Questa finestra di dialogo può essere aperta dalla finestra di dialogo Applicazioni di supporto e dalle finestre di dialogo di creazione, visualizzazione e applicazione dei nodi correlati a ODM.

Figura 4-32
Finestra di dialogo Oracle List Models (Elenco modelli Oracle)



Di seguito sono riportate le informazioni visualizzate per ogni modello:

- **Nome modello.** Nome del modello, che viene utilizzato per ordinare l'elenco
- **Informazioni sul modello.** Informazioni chiave sul modello, inclusive di data e ora di creazione e nome della colonna obiettivo
- **Tipo di modello.** Nome dell'algoritmo di creazione del modello

Oracle Data Miner

Oracle Data Miner è l'interfaccia utente relativa a Oracle Data Mining (ODM) e sostituisce l'interfaccia utente IBM® SPSS® Modeler precedente. Oracle Data Miner è stato appositamente progettato per incrementare l'utilizzo corretto degli algoritmi ODM da parte degli analisti. Questi obiettivi vengono affrontati e risolti in diversi modi:

- Gli utenti necessitano di maggiore assistenza nell'applicazione di una metodologia che gestisce sia la preparazione dei dati sia la selezione degli algoritmi. Oracle Data Miner soddisfa questa esigenza fornendo delle specifiche attività di data mining che guidano gli utenti passo passo nell'utilizzo della metodologia corretta.
- Oracle Data Miner contiene un'euristica migliorata e ampliata nelle procedure guidate di creazione e trasformazione dei modelli, che consente di ridurre la possibilità di errori nella specifica delle impostazioni di trasformazione e di modello.

Definizione di una connessione Oracle Data Miner

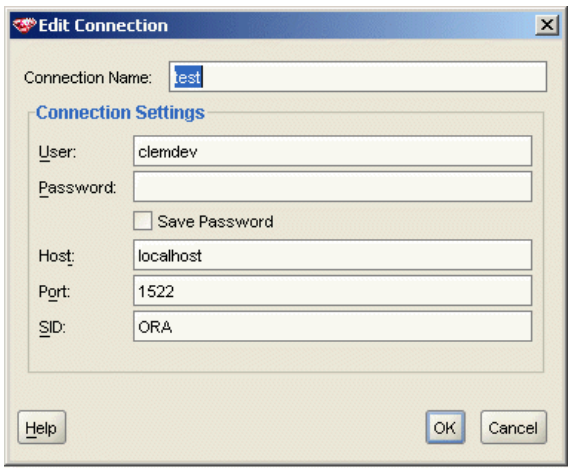
▶ Oracle Data Miner può essere avviato da tutti i nodi Applicazione e Creazione di Oracle e da qualsiasi finestra di dialogo di output mediante il pulsante **Avvia Oracle Data Miner**.

Figura 4-33
Pulsante Avvia Oracle Data Miner

▶ La finestra di dialogo **Edit Connection** di Oracle Data Miner viene visualizzata all'utente prima che venga avviata l'applicazione esterna (a condizione che l'opzione Applicazione di supporto sia stata correttamente definita).

Nota: questa finestra di dialogo viene visualizzata solo in mancanza di un nome di connessione definito.

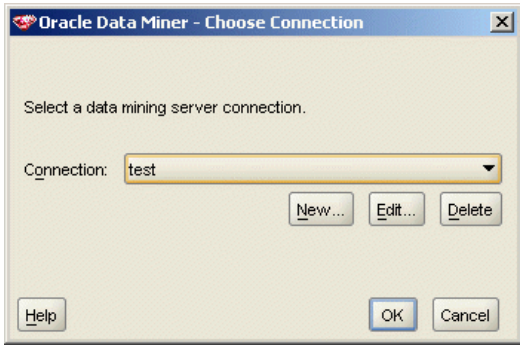
Figura 4-34
Finestra di dialogo Edit Connection di Oracle Data Miner



- Indicare un nome per la connessione Data Miner e immettere le informazioni relative al server Oracle 10gR1 o 10gR2. Il server Oracle dovrebbe essere lo stesso server specificato in SPSS Modeler.

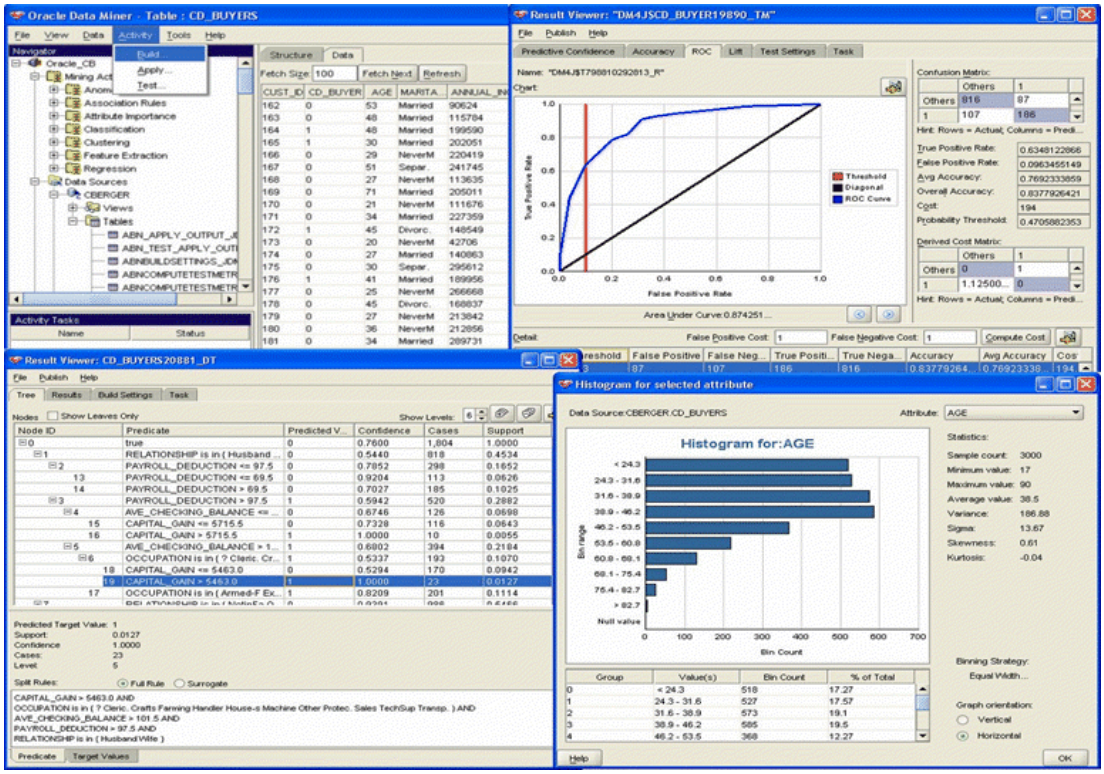
- La finestra di dialogo **Choose Connection** di Oracle Data Miner fornisce le opzioni per specificare il nome della connessione utilizzato, definito al punto precedente.

Figura 4-35
Finestra di dialogo Choose Connection di Oracle Data Miner



Per ulteriori informazioni sui requisiti, l'installazione e l'utilizzo di Oracle Data Miner, consultare la sezione Oracle Data Miner (*http://www.oracle.com/technology/products/bi/odm/odminer/odminer_install_102.htm*) sul sito Web di Oracle.

Figura 4-36
Interfaccia utente Oracle Data Miner



Preparazione dei dati

Quando per la modellazione si utilizzano i modelli Bayes naive, Bayes adattivo e SVM forniti con gli algoritmi Oracle Data Mining possono essere utili due tipi di preparazione dei dati:

- **Discretizzazione**, o conversione di campi di intervalli numerici continui in categorie per algoritmi che non possono accettare dati continui.
- **Normalizzazione**, o trasformazioni applicate a intervalli numerici in modo che abbiano medie e deviazioni standard simili.

Categorizzazione

Il nodo Discretizza di IBM® SPSS® Modeler offre diverse tecniche per l'esecuzione di operazioni di discretizzazione. Viene definita un'operazione di discretizzazione applicabile a uno o più campi. Se si esegue l'operazione di discretizzazione su un insieme di dati vengono create le soglie e consentita la creazione del nodo Nuovo campo di SPSS Modeler. L'operazione Nuovo campo può essere convertita in SQL e applicata prima della creazione e del calcolo del punteggio del modello. Questo approccio crea una dipendenza tra il modello e il nodo Nuovo campo che esegue la discretizzazione, ma consente di riutilizzare le specifiche di discretizzazione in diverse operazioni di modellazione.

Normalizzazione

I campi continui (intervallo numerico) utilizzati come input dei modelli SVM devono essere normalizzati prima della creazione del modello. Nel cado di modelli di regressione, la normalizzazione deve anche essere invertita per ricreare il punteggio dall'output del modello. Tra le impostazioni del modello SVM è possibile scegliere Punteggio Z, Min-Max o Nessuno. I coefficienti di normalizzazione vengono creati da Oracle durante il processo di creazione del modello, per poi essere caricati in SPSS Modeler e archiviati con il modello. In fase di applicazione, i coefficienti vengono convertiti in espressioni di derivazione SPSS Modeler e utilizzati per preparare i dati per il calcolo del punteggio prima che siano passati al modello. In questo caso, la normalizzazione è strettamente associata all'operazione di modellazione.

Esempi di Oracle Data Mining

È disponibile un'ampia gamma di stream di esempio che illustrano l'utilizzo di ODM con IBM® SPSS® Modeler. Tali stream sono disponibili nella cartella di installazione di SPSS Modeler in *\Demos\Database_Modelling\Oracle Data Mining*.

Nota: alla cartella Demos è possibile accedere dal gruppo di programmi SPSS Modeler del menu Start di Windows.

I seguenti stream possono essere utilizzati insieme, in sequenza, come esempio del processo di mining in-database utilizzando l'algoritmo SVM fornito con Oracle Data Mining:

Stream	Descrizione
1_upload_data.str	Utilizzato per la pulitura e il caricamento di dati da un file piatto nel database.
2_explore_data.str	Utilizzato come esempio di esplorazione dei dati con SPSS Modeler.
3_build_model.str	Genera il modello utilizzando l'algoritmo nativo del database.
4_evaluate_model.str	Utilizzato come esempio di valutazione di modelli con SPSS Modeler.
5_deploy_model.str	Esegue il deployment del modello ai fini del calcolo del punteggio in-database.

Nota: Per eseguire l'esempio, gli stream devono essere eseguiti in ordine. Inoltre, i nodi Origine e Modelli in ogni stream devono essere aggiornati per far riferimento a un'origine dati valida per il database che si desidera utilizzare.

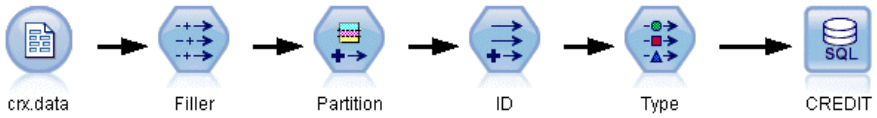
L'insieme di dati impiegato negli stream di esempio concerne applicazioni relative alle carte di credito e presenta un problema di classificazione relativo a un insieme misto di predittori continui e categoriali. Per ulteriori informazioni su questo insieme di dati, vedere il file *crx.names* nella stessa cartella degli stream di esempio.

Questo insieme di dati è disponibile in UCI Machine Learning Repository alla pagina *ftp://ftp.ics.uci.edu/pub/machine-learning-databases/credit-screening/*

Stream di esempio: Caricamento dati

Il primo stream di esempio, *1_upload_data.str*, viene utilizzato per pulire e caricare dati da un file piatto in Oracle.

Figura 4-37
Stream di esempio usato per il caricamento dei dati



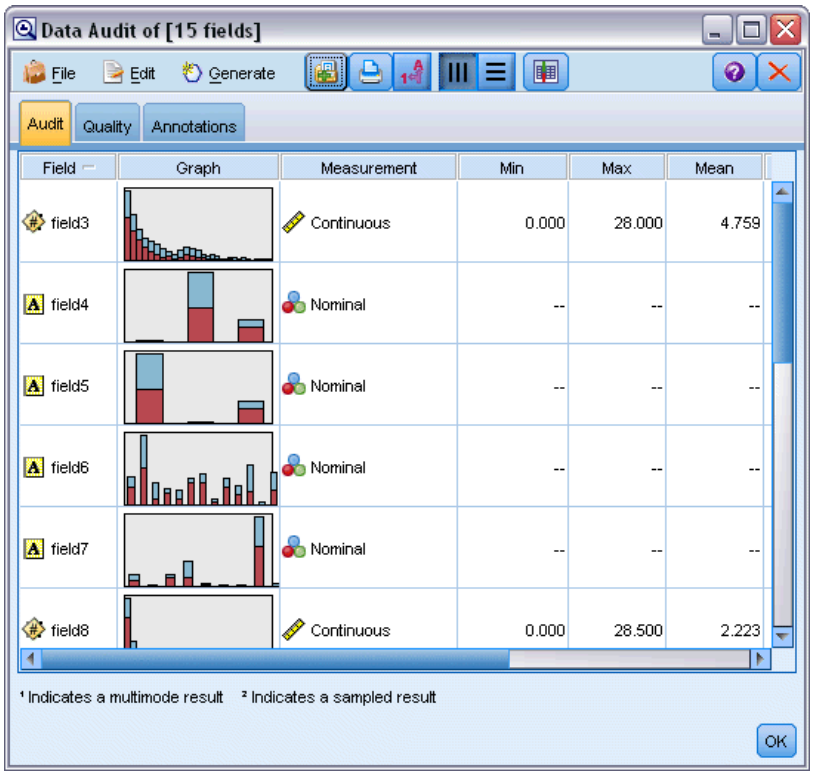
Poiché Oracle Data Mining richiede la specifica di un campo ID univoco, questo stream iniziale utilizza un nodo Nuovo campo per aggiungere all'insieme di dati un nuovo campo denominato *ID* con i valori univoci 1,2,3, mediante la funzione IBM® SPSS® Modeler@INDEX.

Il nodo Riempimento viene utilizzato per la gestione dei valori mancanti e sostituisce i campi vuoti letti dal file di testo *crx.data* con valori *NULLO*.

Stream di esempio: Explore Data

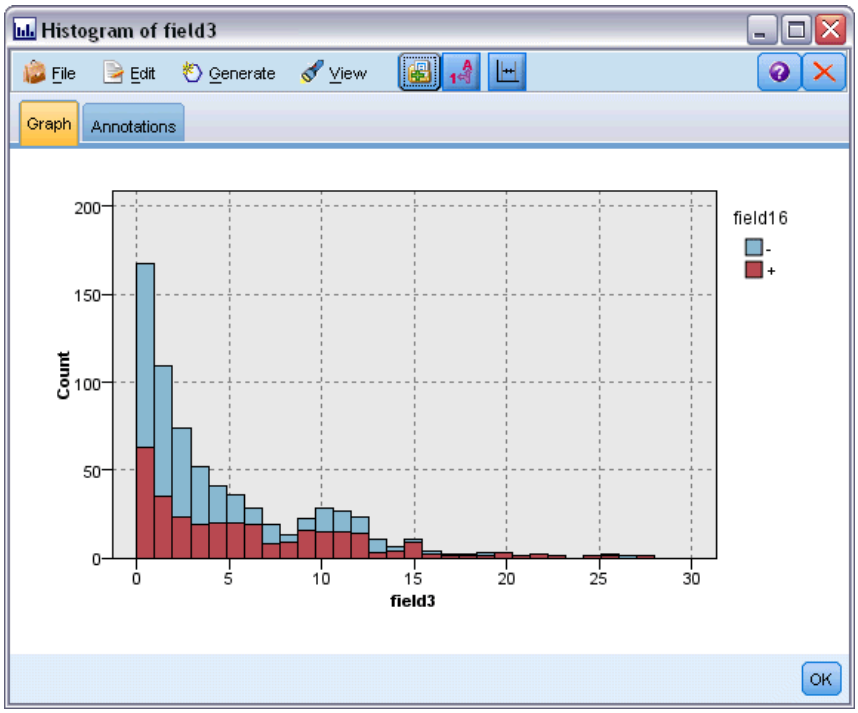
Il secondo stream di esempio, *2_explore_data.str*, viene utilizzato per illustrare l'uso di un nodo Esplora per acquisire una panoramica generale dei dati, comprese statistiche riassuntive e grafici. Per ulteriori informazioni, vedere l'argomento Nodo Esplora in il capitolo 6 in *IBM SPSS Modeler 15 Nodi di input, elaborazione e output*.

Figura 4-38
Risultati di Esplora



Facendo doppio clic su un grafico nel report Esplora, si accede a una visualizzazione più dettagliata del grafico che consente un'esplorazione più approfondita di un dato campo.

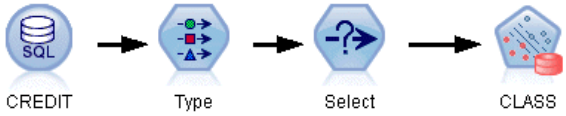
Figura 4-39
Istogramma creato facendo doppio clic su un grafico nella finestra Data Audit



Stream di esempio: Build Model

Il terzo stream di esempio, *3_build_model.str*, illustra la creazione di modelli in IBM® SPSS® Modeler. Fare doppio clic sul nodo di input Database (etichettato CREDIT) per specificare la sorgente di dati. Per specificare le impostazioni di creazione, fare doppio clic sul nodo di creazione (inizialmente etichettato CLASS e modificato in FIELD16 quando viene specificata la sorgente di dati).

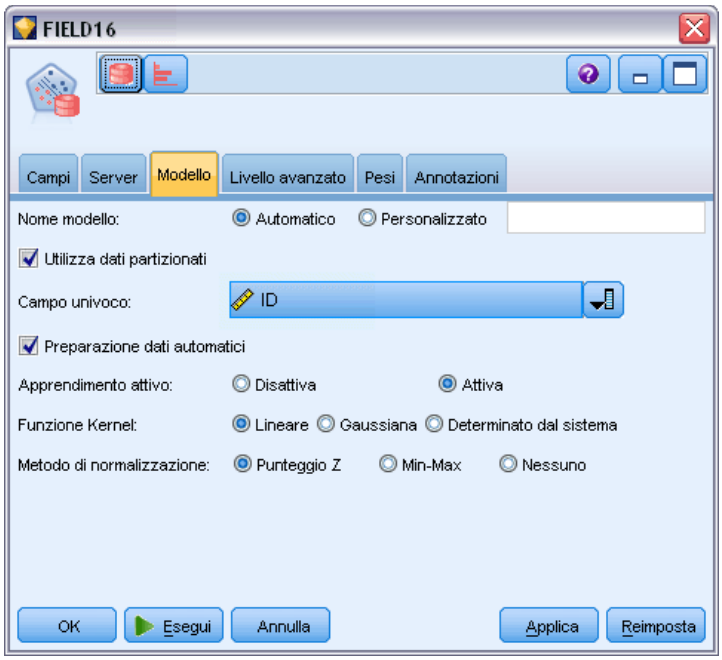
Figura 4-40
Stream di esempio per la modellazione in-database



Nella scheda Modello della finestra di dialogo:

- Verificare che ID sia selezionato come campo Univoco.
- Verificare che Lineare sia selezionato come funzione Kernel e Punteggio Z come metodo di normalizzazione.

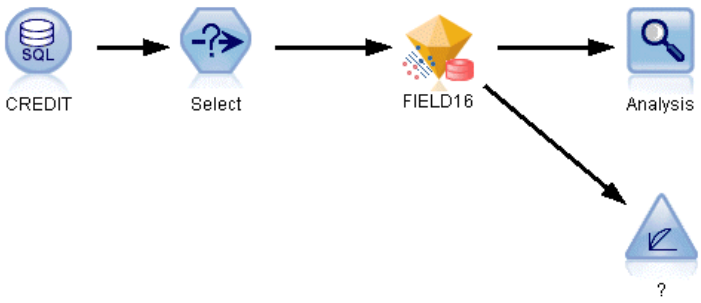
Figura 4-41
Opzioni del modello per SVM Oracle



Stream di esempio: Valutazione modello

Il quarto stream di esempio, *4_evaluate_model.str*, illustra i vantaggi associati all'utilizzo di IBM® SPSS® Modeler per la modellazione in-database. Una volta eseguito il modello, è possibile aggiungerlo nuovamente allo stream di dati e valutarlo con il supporto di un'ampia gamma di strumenti mirati disponibili in SPSS Modeler.

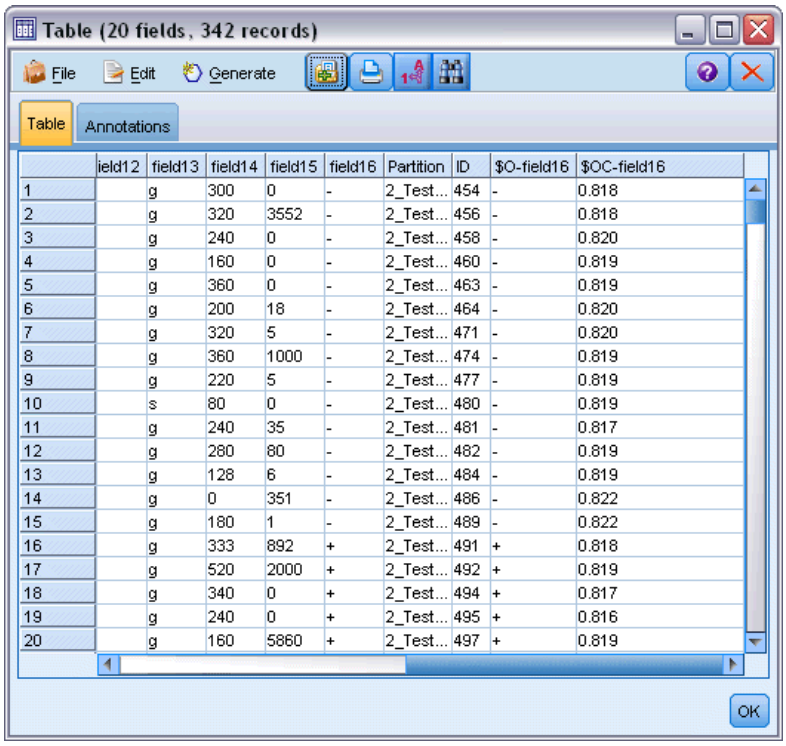
Figura 4-42
Stream di esempio usato per la valutazione del modello



Visualizzazione dei risultati della modellazione

Collegare un nodo Tabella all'insieme di modelli per esplorare i risultati. Il campo $O-field16 mostra il valore previsto per *field16* in ciascun caso e il campo $OC-field16 mostra il valore di confidenza per questa previsione.

Figura 4-43
Tabella con informazioni sulle previsioni generate

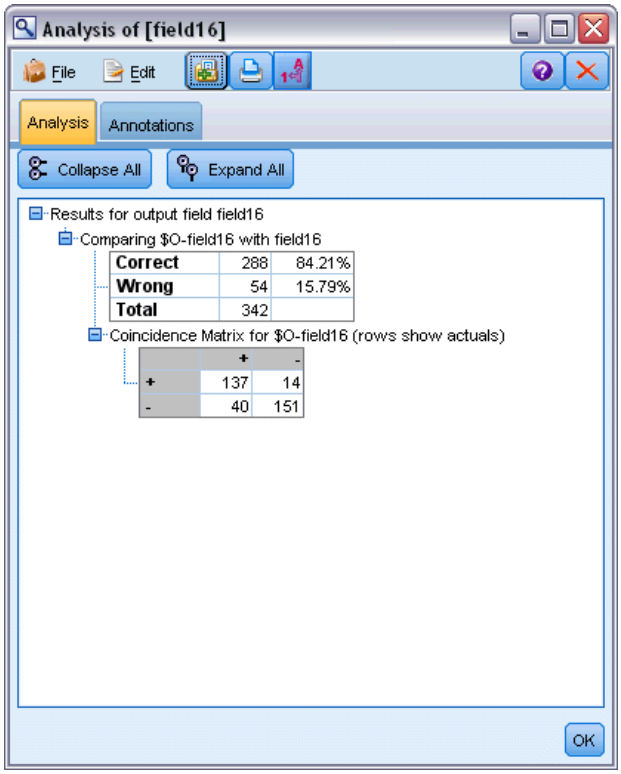


	ield12	field13	field14	field15	field16	Partition	ID	$O-field16	$OC-field16
1		g	300	0	-	2_Test...	454	-	0.818
2		g	320	3552	-	2_Test...	456	-	0.818
3		g	240	0	-	2_Test...	458	-	0.820
4		g	160	0	-	2_Test...	460	-	0.819
5		g	360	0	-	2_Test...	463	-	0.819
6		g	200	18	-	2_Test...	464	-	0.820
7		g	320	5	-	2_Test...	471	-	0.820
8		g	360	1000	-	2_Test...	474	-	0.819
9		g	220	5	-	2_Test...	477	-	0.819
10		s	80	0	-	2_Test...	480	-	0.819
11		g	240	35	-	2_Test...	481	-	0.817
12		g	280	80	-	2_Test...	482	-	0.819
13		g	128	6	-	2_Test...	484	-	0.819
14		g	0	351	-	2_Test...	486	-	0.822
15		g	180	1	-	2_Test...	489	-	0.822
16		g	333	892	+	2_Test...	491	+	0.818
17		g	520	2000	+	2_Test...	492	+	0.819
18		g	340	0	+	2_Test...	494	+	0.817
19		g	240	0	+	2_Test...	495	+	0.816
20		g	160	5860	+	2_Test...	497	+	0.819

Valutazione dei risultati della modellazione

È possibile utilizzare il nodo Analisi per creare una matrice di coincidenza che mostri lo schema di corrispondenze tra ogni campo previsto e il relativo campo obiettivo. Eseguire il nodo Analisi per visualizzare i risultati.

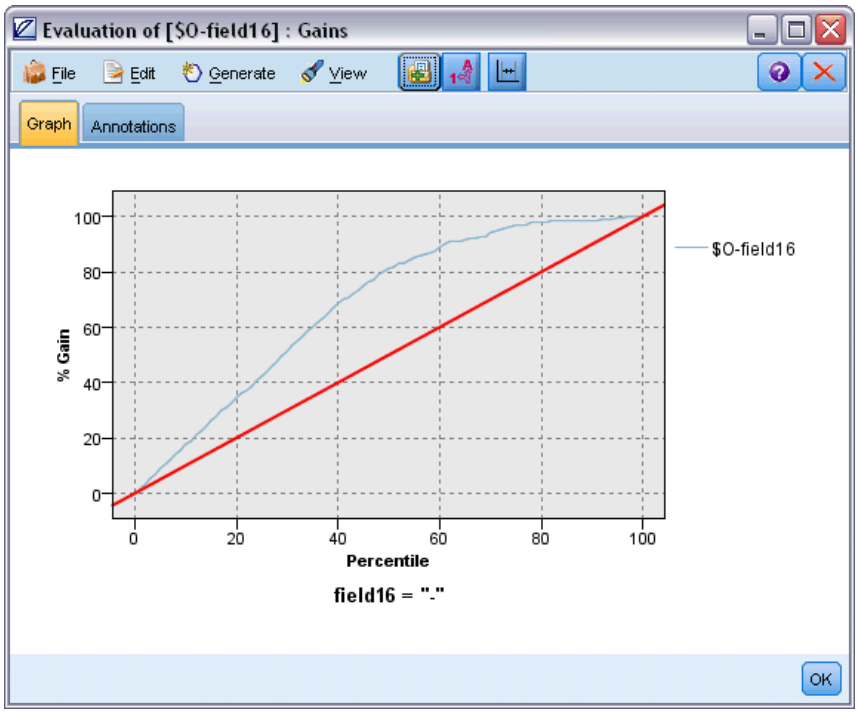
Figura 4-44
Scheda Analisi con informazioni sui risultati di analisi



La tabella indica che l'84.21% delle previsioni create dall'algoritmo SVM Oracle era corretto.

È possibile utilizzare il nodo Valutazione per creare un grafico dei guadagni, progettato per mostrare i miglioramenti in termini di precisione realizzati dal modello. Eseguire il nodo Valutazione per visualizzare i risultati.

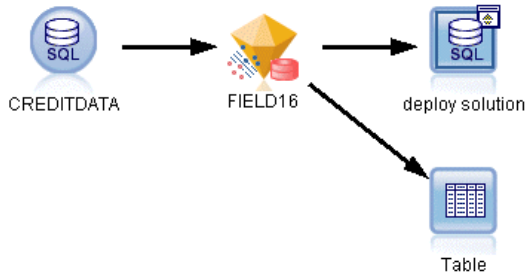
Figura 4-45
Grafico dei guadagni con informazioni sui miglioramenti in termini di precisione realizzati dal modello



Stream di esempio: Deployment modello

Una volta raggiunto il livello di precisione del modello desiderato, è possibile eseguire il deployment del modello per consentirne l'utilizzo con applicazioni esterne o la ripubblicazione nel database. Nell'ultimo stream di esempio, *5_deploy_model.str*, i dati vengono letti dalla tabella CREDITDATA, quindi viene eseguito il calcolo del punteggio e, infine, i dati vengono pubblicati nella tabella CREDITSCORES mediante il nodo Publisher denominato *soluzione di deployment*.

Figura 4-46
Stream di esempio per la modellazione in-database



Per ulteriori informazioni, vedere l'argomento Funzionamento di IBM SPSS Modeler Solution Publisher in il capitolo 2 in *IBM SPSS Modeler 15 Solution Publisher*.

Capitolo
5

Modellazione di database con IBM InfoSphere Warehouse

IBM InfoSphere Warehouse e IBM SPSS Modeler

IBM InfoSphere Warehouse (ISW) fornisce una famiglia di algoritmi di data mining incorporati nel sistema DB2 RDBMS di IBM. IBM® SPSS® Modeler è dotato di nodi che supportano l'integrazione dei seguenti algoritmi IBM:

- Alberi decisionali
- Regole di associazione
- Raggruppamento tramite cluster demografici
- Raggruppamento tramite cluster Kohonen
- Regole di sequenza
- Regressione trasformazione
- Regressione lineare
- Regressione polinomiale
- Bayes naive
- Regressione logistica
- Serie storiche

Per ulteriori informazioni su questi algoritmi, consultare la documentazione fornita con l'installazione di IBM InfoSphere Warehouse.

Requisiti per l'integrazione con IBM InfoSphere Warehouse

Di seguito sono elencate le condizioni che costituiscono i prerequisiti indispensabili per l'esecuzione della modellazione in-database con InfoSphere Warehouse Data Mining. Per garantire che queste condizioni vengano soddisfatte, può essere necessario consultare l'amministratore di sistema.

- Esecuzione di IBM® SPSS® Modeler in un'installazione di IBM® SPSS® Modeler Server su Windows o UNIX.
- IBM DB2 Data Warehouse Edition Versione 9.1

  o
- IBM InfoSphere Warehouse Versione 9.5 Enterprise Edition
- Una sorgente dati ODBC per la connessione a DB2, come illustrato di seguito.

Nota: le funzionalità di modellazione in-database e ottimizzazione SQL richiedono che sul computer SPSS Modeler sia attivata la connettività SPSS Modeler Server. Con questa impostazione attivata, è possibile accedere agli algoritmi di database, restituire codice SQL direttamente da SPSS Modeler e accedere a SPSS Modeler Server. Per verificare lo stato attuale della licenza, scegliere le seguenti opzioni dal menu SPSS Modeler.
Guida > Informazioni su > Ulteriori dettagli

Se la connettività è abilitata, l'opzioneAbilitazione server viene visualizzata nella scheda Stato della licenza.

Per ulteriori informazioni, vedere l'argomento Connessione a IBM SPSS Modeler Server in il capitolo 3 in *Manuale dell'utente di IBM SPSS Modeler 15*.

Attivazione dell'integrazione con IBM InfoSphere Warehouse

Per attivare l'integrazione di IBM® SPSS® Modeler con IBM InfoSphere Warehouse (ISW) Data Mining, sarà necessario configurare ISW e creare una sorgente ODBC, attivare l'integrazione nella finestra di dialogo Applicazioni di supporto di SPSS Modeler e abilitare la generazione e l'ottimizzazione SQL.

Configurazione di ISW

Per installare e configurare ISW seguire le istruzioni contenute nella guida all'installazione di *InfoSphere Warehouse*.

Creazione di una sorgente ODBC per ISW

Per attivare la connessione tra ISW e SPSS Modeler è necessario creare un nome di sorgente dati (DSN, Data Source Name) ODBC.

Per creare un DSN, è necessario avere una conoscenza di base delle sorgenti dati e dei driver ODBC e disporre del supporto database in SPSS Modeler. Per ulteriori informazioni, vedere l'argomento Accesso ai dati in il capitolo 2 in *IBM SPSS Modeler Server 15 Guida della performance e amministrazione*.

Se IBM® SPSS® Modeler Server e IBM InfoSphere Warehouse Data Mining sono in esecuzione su host differenti, creare lo stesso DSN ODBC su ogni host. Assicurarsi di utilizzare lo stesso nome per il DSN su entrambi gli host.

- Installare i driver ODBC. I driver sono disponibili sul disco di installazione di IBM® SPSS® Data Access Pack fornito con questa versione. Eseguire il file *setup.exe* per avviare il programma di installazione, e selezionare tutti i driver opportuni. Attenersi alle istruzioni visualizzate per installare i driver.
- Creare il DSN.

  Nota: la sequenza dei menu dipende dalla versione di Windows in uso.

  - **Windows XP.** Dal menu Start, scegliere Pannello di controllo. Fare doppio clic su Strumenti di amministrazione, quindi ancora doppio clic su Origini dati (ODBC).

- **Windows Vista.** Dal menu Start, scegliere Pannello di controllo, quindi Strumenti di amministrazione. Fare doppio clic su Strumenti di amministrazione, selezionare Origini dati (ODBC) quindi Apri.
- **In Windows 7.** Dal menu Start, scegliere Pannello di controllo, quindi Sistema e sicurezza e Strumenti di amministrazione. Selezionare Origini dati (ODBC) e fare clic su Apri.

▶ Fare clic sulla scheda DSN di sistema, quindi fare clic su Aggiungi.

▶ Selezionare il driver SPSS OEM 6.0 DB2 Wire Protocol.

▶ Fare clic su Fine.

▶ Nella finestra di dialogo ODBC DB2 Wire Protocol Driver Setup:

- Specificare un nome di sorgente dati.
- Per l'indirizzo IP, indicare il nome host del server su cui è in esecuzione il sistema DB2 RDBMS.
- Accettare il valore di default relativo alla porta TCP (50000).
- Specificare il nome del database con il quale verrà stabilita la connessione.

▶ Fare clic su Verifica connessione.

▶ Nella finestra di dialogo Connessione a DB2 Wire Protocol immettere il nome utente e la password ricevuti dall'amministratore di sistema, quindi fare clic su OK.

Verrà visualizzato il messaggio Connessione effettuata.

Driver IBM DB2 ODBC. Se il driver ODBC in uso corrisponde al driver IBM DB2 ODBC, applicare la seguente procedura per creare un DSN ODBC:

▶ In Amministratore origine dati ODBC fare clic sulla scheda DSN di sistema, quindi fare clic su Aggiungi.

▶ Selezionare IBM DB2 ODBC DRIVER, quindi fare clic su Fine.

▶ Nella finestra IBM DB2 ODBC DRIVER—Add, immettere un nome di sorgente dati, quindi per l'alias di database fare clic su Aggiungi.

▶ Nella finestra CLI/ODBC Settings—<Data source name>, all'interno della scheda Origine dati, immettere l'ID utente e la password ricevuti dall'amministratore di sistema, quindi fare clic sulla scheda TCP/IP.

▶ Nella scheda TCP/IP immettere:

- Il nome del database al quale si desidera connettersi
- Un nome alias di database (non più di otto caratteri)
- Il nome host del server di database al quale si desidera connettersi
- Il numero di porta per la connessione

▶ Fare clic sulla scheda Opzioni di protezione e selezionare Specifica le opzioni di sicurezza (Opzionale), quindi accettare l'impostazione di default (Utilizza il valore di autenticazione nella configurazione DBM del server).

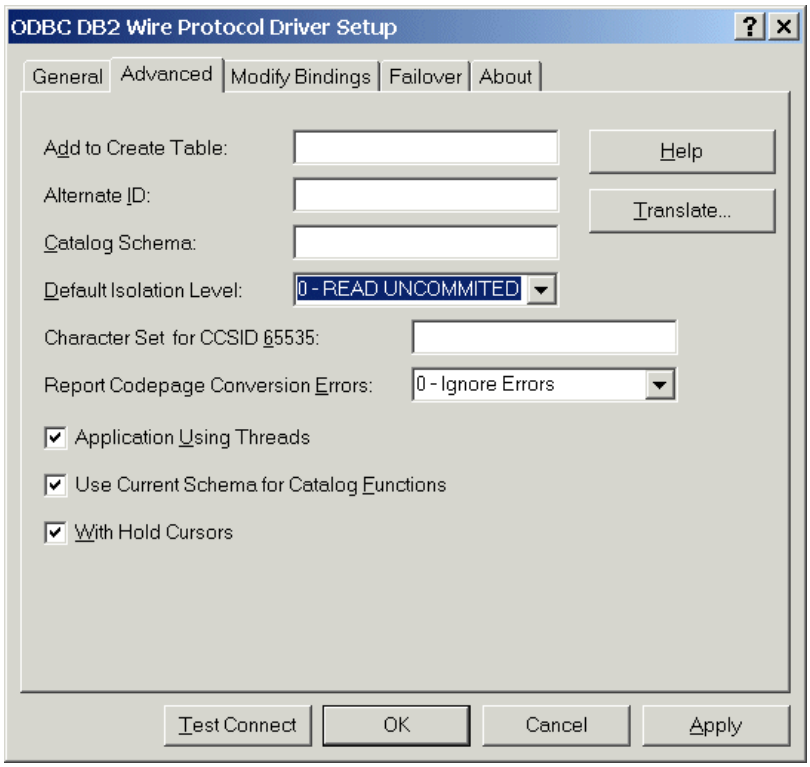
▶ Fare clic sulla scheda Origine dati, quindi su Connetti.

Verrà visualizzato il messaggio Connessione verificata.

Configurazione di ODBC per il feedback (opzionale)

Per ricevere feedback da IBM InfoSphere Warehouse Data Mining durante la creazione di modelli e attivare SPSS Modeler per l'annullamento del processo di creazione, attenersi alla procedura riportata di seguito in modo da configurare la sorgente dati ODBC creata nella sezione precedente. Si noti che questa procedura di configurazione consente a SPSS Modeler di leggere i dati DB2 che non possono essere salvati nel database eseguendo contemporaneamente transazioni. Se si nutrono dubbi circa le implicazioni di questa modifica, si consiglia di consultare l'amministratore di sistema.

Figura 5-1
Finestra di dialogo ODBC DB2 Wire Protocol Driver Setup, scheda Avanzate



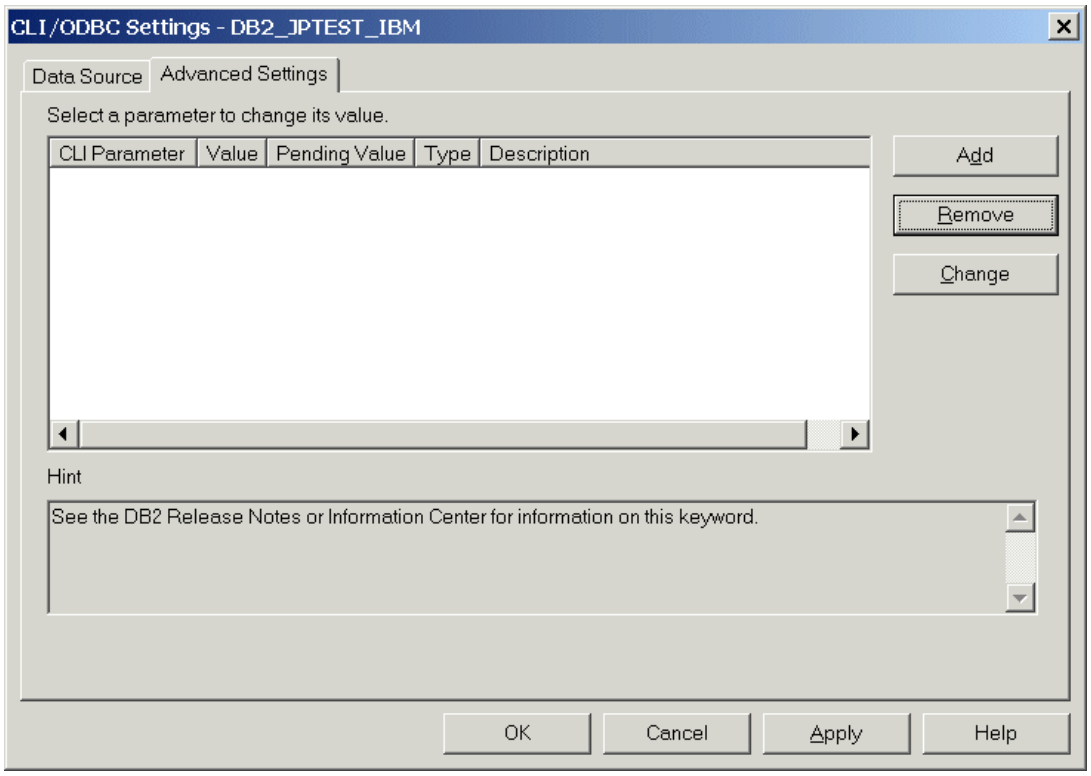
Driver SPSS OEM 6.0 DB2 Wire Protocol. Per il driver Connect ODBC, effettuare le seguenti operazioni:

▶ Avviare Amministratore origine dati ODBC, selezionare la sorgente dati creata nella sezione precedente, quindi fare clic sul pulsante Configura.

▶ Nella finestra di dialogo ODBC DB2 Wire Protocol Driver Setup fare clic sulla scheda Avanzate.

▶ Impostare il livello di isolamento di default su 0-READ UNCOMMITED, quindi fare clic su OK.

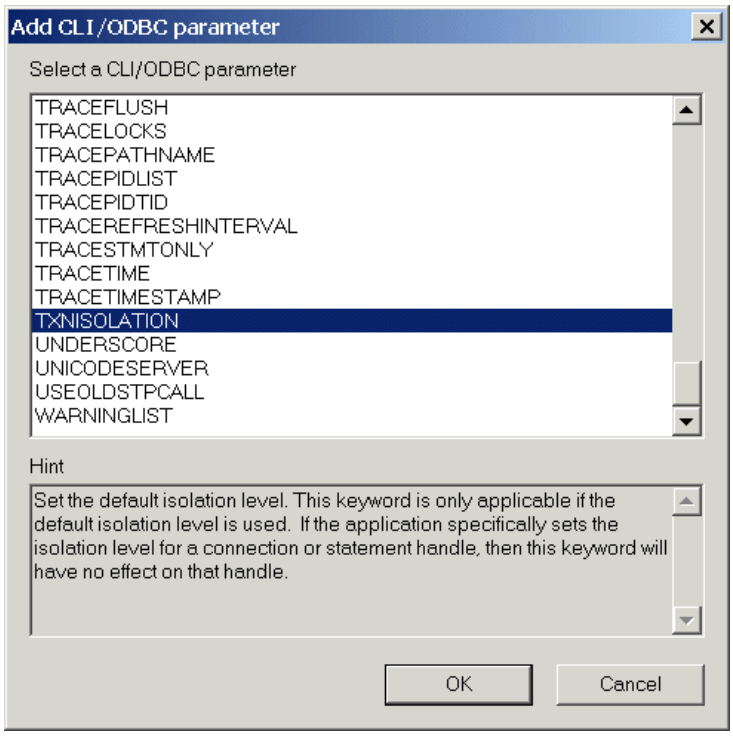
Figura 5-2
Finestra di dialogo Impostazioni CLI/ODBC, scheda Impostazioni avanzate



Driver IBM DB2 ODBC. Per il driver IBM DB2, effettuare le seguenti operazioni:

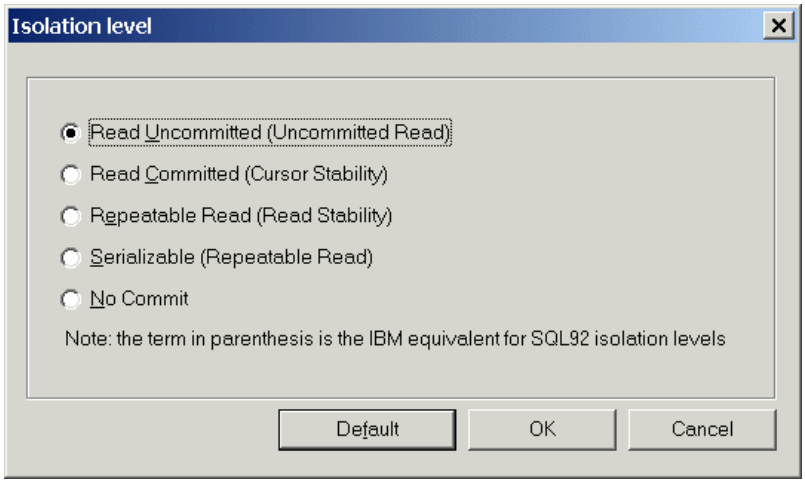
- Avviare Amministratore origine dati ODBC, selezionare la sorgente dati creata nella sezione precedente, quindi fare clic sul pulsante Configura.
- Nella finestra di dialogo CLI/ODBC Settings fare clic sulla scheda Impostazioni avanzate, quindi sul pulsante Aggiungi.

Figura 5-3
Finestra di dialogo CLI/ODBC Parameter



- Nella finestra di dialogo Add CLI/ODBC Parameter selezionare il parametro TXNISOLATION, quindi fare clic su OK.

Figura 5-4
Finestra di dialogo Isolation level



- Nella finestra di dialogo Isolation level selezionare Read Uncommitted, quindi fare clic su OK.

- Nella finestra di dialogo CLI/ODBC Settings, fare clic su OK per concludere la configurazione.

Si noti che il feedback riportato da IBM InfoSphere Warehouse Data Mining viene visualizzato nel seguente formato:

<ITERATIONNO> / <PROGRESS> / <KERNELPHASE>

dove:

- <ITERATIONNO> indica il numero del passaggio corrente sui dati, a partire da 1.
- <PROGRESS> indica lo stato di avanzamento dell'iterazione corrente sotto forma di un numero compreso tra 0.0 e 1.0.
- <KERNELPHASE> descrive la fase corrente dell'algoritmo di data mining.

Attivazione di IBM InfoSphere Warehouse Data Mining Integration in IBM SPSS Modeler

Per attivare SPSS Modeler in modo da consentire l'utilizzo di DB2 con IBM InfoSphere Warehouse Data Mining, è necessario prima fornire alcune specifiche nella finestra di dialogo Applicazioni di supporto.

▶ Dai menu di SPSS Modeler scegliere:
Strumenti > Opzioni > Applicazioni di supporto

▶ Fare clic sulla scheda IBM InfoSphere Warehouse.

Attiva InfoSphere Warehouse Data Mining Integration. Attiva la palette Modelli in-database (se non è già visualizzata) nella parte inferiore della finestra SPSS Modeler e aggiunge i nodi degli algoritmi di ISW Data Mining.

Connessione DB2. Specifica la sorgente dati ODBC DB2 di default utilizzata per la creazione e l'archiviazione di modelli. Questa impostazione può essere sovrascritta nei singoli nodi di modelli generati e creazione modelli. Fare clic sul pulsante con i puntini di sospensione (...) per scegliere la sorgente dati.

la connessione al database utilizzata a fini di modellazione può corrispondere o meno a quella impiegata per accedere ai dati. Per esempio, è possibile utilizzare uno stream che accede ai dati di un database DB2, li scarica in SPSS Modeler per la pulitura o altre operazioni di modifica e, infine, li carica su un database DB2 differente per la modellazione. In alternativa, i dati originali possono risiedere in un file piatto o in un'altra sorgente (non DB2) e occorrerà pertanto caricarli in DB2 per il processo di modellazione. In tutti i casi, se necessario, i dati verranno automaticamente caricati in una tabella temporanea creata nel database utilizzato per la modellazione.

Warn when about to overwrite an InfoSphere Warehouse Data Mining Integration model. Selezionare questa opzione per assicurarsi che i modelli archiviati nel database non vengano sovrascritti da SPSS Modeler senza preavviso.

Elenca modelli di InfoSphere Warehouse Data Mining. Consente di elencare ed eliminare i modelli archiviati in DB2. Per ulteriori informazioni, vedere l'argomento Elenco dei modelli in-database a pag. 116.

Attiva lancio della visualizzazione di InfoSphere Warehouse Data Mining. Se è stato installato il modulo Visualization, è necessario attivarlo qui per l'utilizzo di SPSS Modeler.

Percorso file eseguibile di Visualization. La posizione del file eseguibile del modulo Visualization (se installato), per esempio *C:\Programmi\IBM\ISWarehouse\Im\IMVisualization\bin\imvisualizer.exe*.

Directory plug-in di visualizzazione serie storica. Il percorso del plug-in Flash di visualizzazione delle serie storiche (se installato), per esempio *C:\Programmi\IBM\ISWShared\plugins\com.ibm.datatools.datamining.imvisualization.flash_2.2.1.v20091111_0915*

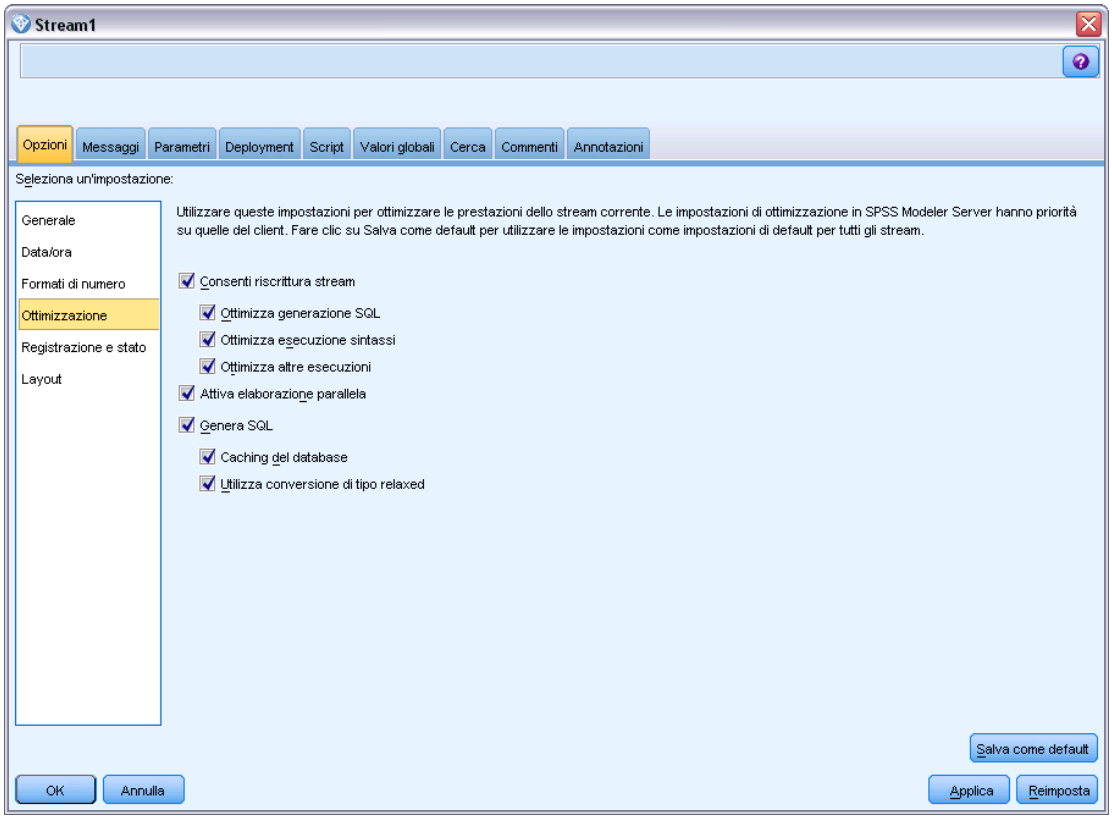
Attiva opzioni avanzate di InfoSphere Warehouse Data Mining. È possibile impostare un limite all'utilizzo di memoria su un algoritmo di mining in-database e specificare altre opzioni arbitrarie sotto forma di riga di comando per modelli specifici. La definizione del limite consente di controllare l'utilizzo di memoria e specificare un valore per l'opzione avanzata `-buf`. È possibile specificare in questa posizione anche altre opzioni avanzate sotto forma di riga di comando e passarle a IBM InfoSphere Warehouse Data Mining. Per ulteriori informazioni, vedere l'argomento Opzioni avanzate a pag. 120.

Controlla versione di InfoSphere Warehouse. Controlla la versione di IBM InfoSphere Warehouse in uso e visualizza un messaggio di errore se si tenta di utilizzare una funzione di data mining non supportata in quella versione.

Attivazione di generazione e ottimizzazione SQL

- Dai menu di SPSS Modeler scegliere:
  Strumenti > Proprietà stream > Opzioni

Figura 5-5
Impostazioni di ottimizzazione



- Fare clic sull'opzione Ottimizzazione nel riquadro di spostamento.

- Confermare che l'opzione Genera SQL è attivata. Questa impostazione è necessaria per il corretto funzionamento della modellazione di database.
- Selezionare Ottimizza generazione SQL e Ottimizza altre esecuzioni (queste due opzioni non sono strettamente necessarie, tuttavia se ne consiglia la selezione per ottenere performance ottimizzate).

Per ulteriori informazioni, vedere l'argomento Impostazione delle opzioni di ottimizzazione per gli stream in il capitolo 5 in *Manuale dell'utente di IBM SPSS Modeler 15*.

Creazione di modelli con IBM InfoSphere Warehouse Data Mining

La creazione di modelli di IBM InfoSphere Warehouse Data Mining richiede che l'insieme di dati addestramento sia posizionato in una tabella o visualizzazione all'interno del database DB2. Se i dati non sono ubicati in DB2 o devono essere elaborati in IBM® SPSS® Modeler come parte del processo di preparazione dei dati che non è possibile eseguire in DB2, tali dati vengono automaticamente caricati in una tabella temporanea di DB2 prima della creazione dei modelli.

Calcolo del punteggio e deployment dei modelli

Il calcolo del punteggio dei modelli avviene sempre all'interno di DB2 ed è sempre eseguito da IBM InfoSphere Warehouse Data Mining. Può essere necessario caricare l'insieme di dati in una tabella temporanea, qualora i dati vengano originati in IBM® SPSS® Modeler o debbano essere preparati all'interno dell'applicazione. In SPSS Modeler, per i modelli di cluster, Albero decisionale e Regressione viene generalmente fornita una sola previsione con la probabilità o la confidenza associata. È inoltre disponibile un'opzione utente per la visualizzazione delle confidenze per ogni possibile risultato (simile a quella della regressione logistica) che rappresenta un'opzione tempo punteggio ubicata all'interno della scheda Impostazioni dell'insieme di modelli (la casella di controllo Includi confidenze per tutte le classi). Per i modelli di sequenza e associazione di SPSS Modeler vengono forniti diversi valori. SPSS Modeler può calcolare i modelli IBM InfoSphere Warehouse Data Mining da stream pubblicati per l'esecuzione mediante IBM® SPSS® Modeler Solution Publisher.

I seguenti campi sono generati da modelli di calcolo del punteggio:

Tabella 5-1
Campi di calcolo del punteggio dei modelli

Tipo di modello	Colonne punteggio	Significato
Alberi decisionali	$I-*campo*	Previsione migliore per il *campo*.
	$IC-*campo*	Confidenza della previsione migliore per il *campo*.
	$IC-*valore1*, ..., $IC-*valoreN*	(facoltativo) Confidenza di ciascuno dei possibili valori *N* per il *campo*.
Regression	$I-*campo*	Previsione migliore per il *campo*.
	$IC-*campo*	Confidenza della previsione migliore per il *campo*.

Tipo di modello	Colonne punteggio	Significato
Raggruppamento tramite cluster	$I-*nome_modello*	Migliore assegnazione cluster per il record di input.
	$IC-*nome_modello*	Confidenza della migliore assegnazione cluster per il record di input.
Associazione	$I-*nome_modello*	Identificatore della regola corrispondente.
	$IH-*nome_modello*	Elemento dell'intestazione.
	$IHN-*nome_modello*	Nome dell'elemento dell'intestazione.
	$IS-*nome_modello*	Valore di supporto della regola corrispondente.
	$IC-*nome_modello*	Valore di confidenza della regola corrispondente.
	$IL-*nome_modello*	Valore di lift della regola corrispondente.
	$IMB-*nome_modello*	Numero di elementi del corpo o di insiemi di elementi del corpo corrispondenti (dal momento che tutti gli elementi o gli insiemi di elementi del corpo devono corrispondere a questo numero, esso è uguale al numero di elementi o di insiemi di elementi del corpo).
Sequenza	$I-*nome_modello*	Identificatore della regola corrispondente
	$IH-*nome_modello*	Insieme di elementi dell'intestazione della regola corrispondente
	$IHN-*nome_modello*	Nomi degli elementi dell'insieme di elementi dell'intestazione della regola corrispondente
	$IS-*nome_modello*	Valore di supporto della regola corrispondente
	$IC-*nome_modello*	Valore di confidenza della regola corrispondente
	$IL-*nome_modello*	Valore di lift della regola corrispondente
	$IMB-*nome_modello*	Numero di elementi del corpo o di insiemi di elementi del corpo corrispondenti (dal momento che tutti gli elementi o gli insiemi di elementi del corpo devono corrispondere a questo numero, esso è uguale al numero di elementi o di insiemi di elementi del corpo)

Tipo di modello	Colonne punteggio	Significato
Bayes naive	$I-*campo*	Previsione migliore per il *campo*.
	$IC-*campo*	Confidenza della previsione migliore per il *campo*.
Regressione logistica	$I-*campo*	Previsione migliore per il *campo*.
	$IC-*campo*	Confidenza della previsione migliore per il *campo*.

Gestione dei modelli DB2

La creazione di un modello di IBM InfoSphere Warehouse Data Mining tramite IBM® SPSS® Modeler comporta la creazione di un modello in SPSS Modeler e la creazione o la sostituzione di un modello nel database DB2. Un modello di SPSS Modeler di questo tipo fa riferimento al contenuto di un modello di database archiviato in un server di database. SPSS Modeler consente di eseguire un controllo dell'uniformità archiviando una stringa identica con la chiave del modello generato sia nel modello DB2 che nel modello di SPSS Modeler.

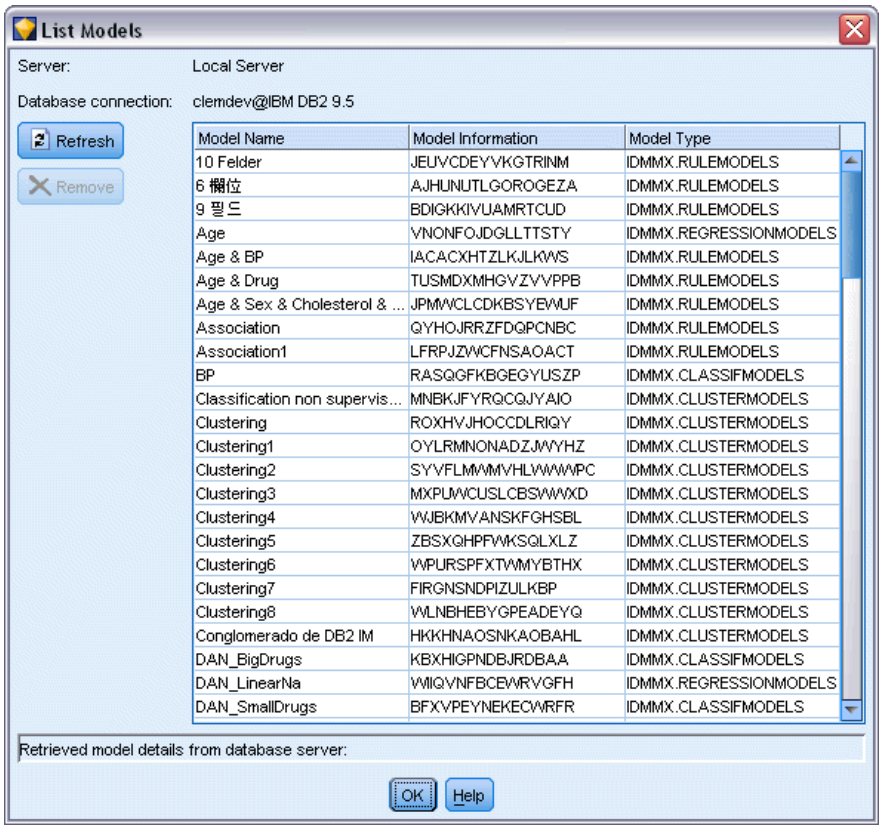
La stringa della chiave di ogni modello DB2 viene visualizzata nella colonna *Informazioni sul modello* all'interno della finestra di dialogo Elenco dei modelli in-database. La stringa della chiave di un modello di SPSS Modeler viene visualizzata come Chiave di modello nella scheda Server di un modello di SPSS Modeler (se all'interno di uno stream).

È possibile utilizzare il pulsante Controllo per verificare la corrispondenza delle chiavi nel modello DB2 e in quello di SPSS Modeler. Se in DB2 non è reperibile alcun modello con lo stesso nome o se le chiavi del modello non corrispondono, il modello DB2 è stato eliminato o ricreato dopo la creazione del modello di SPSS Modeler. Per ulteriori informazioni, vedere l'argomento Scheda Server dell'insieme di modelli ISW a pag. 153.

Elenco dei modelli in-database

IBM® SPSS® Modeler dispone di una finestra di dialogo in cui è possibile elencare i modelli archiviati in IBM InfoSphere Warehouse Data Mining ed eliminare modelli specifici.

Figura 5-6
Finestra di dialogo Elenco dei modelli DB2



Tale finestra di dialogo è accessibile dalla finestra Applicazioni di servizio IBM nonché dalle finestre relative alle operazioni di creazione, visualizzazione e applicazione per i nodi correlati a IBM InfoSphere Warehouse Data Mining. Di seguito sono riportate le informazioni visualizzate per ogni modello:

- Nome del modello (utilizzato per ordinare l'elenco).
- Informazioni sul modello (informazioni sulla chiave di modello, da una chiave casuale che viene generata quando SPSS Modeler crea il modello).
- Tipo di modello (la tabella DB2 in cui IBM InfoSphere Warehouse Data Mining ha archiviato il modello).

Visualizzazione dei modelli

Lo strumento visualizzatore è l'unico modo per sfogliare i modelli di InfoSphere Warehouse Data Mining. Questo strumento può essere installato in via facoltativa con InfoSphere Warehouse Data Mining. Per ulteriori informazioni, vedere l'argomento Attivazione dell'integrazione con IBM InfoSphere Warehouse a pag. 107.

- Fare clic su Visualizza per avviare lo strumento visualizzatore. Ciò che viene visualizzato dallo strumento dipende dal tipo di nodo generato. Per esempio, lo strumento visualizzatore restituirà una visualizzazione Classi previste quando viene avviato da un insieme di modelli Albero decisionale ISW.
- Fare clic su Risultati del test (solo Alberi decisionali e Sequenza) per avviare lo strumento visualizzatore e visualizzare la qualità globale del modello generato.

Esportazione di modelli e generazione di nodi

È possibile eseguire importazioni ed esportazioni del PMML nei modelli di IBM InfoSphere Warehouse Data Mining. Il PMML che viene esportato è quello originale generato da IBM InfoSphere Warehouse Data Mining. La funzione di esportazione restituisce il modello in formato PMML.

È possibile esportare il riepilogo e la struttura di un modello in file formato testo e HTML, nonché generare i nodi Filtro, Seleziona e Nuovo Campo appropriati laddove necessario. Per ulteriori informazioni, vedere "Esportazione di modelli" nel *Manuale dell'utente di IBM® SPSS® Modeler*.

Impostazioni dei nodi comuni a tutti gli algoritmi

Le seguenti impostazioni sono valide per molti degli algoritmi di IBM InfoSphere Warehouse Data Mining:

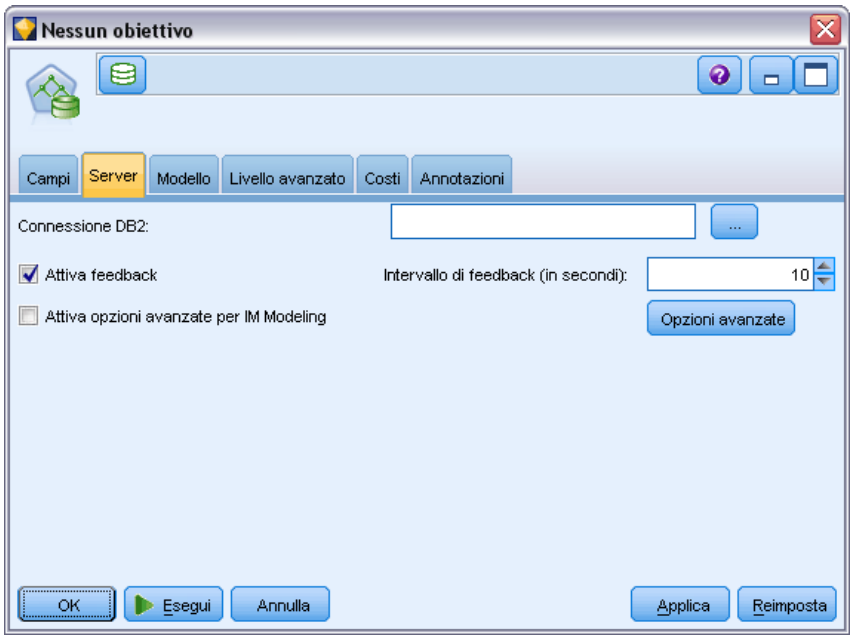
Obiettivo e predittori. È possibile specificare obiettivo e predittori utilizzando il nodo Tipo oppure manualmente mediante la scheda Campi del nodo per la creazione di modelli, come è prassi standard in IBM® SPSS® Modeler.

sorgenti dati ODBC. Consente all'utente di sovrascrivere la sorgente dati ODBC di default per il modello corrente. (L'impostazione di default è specificata nella finestra di dialogo Applicazioni di supporto. Per ulteriori informazioni, vedere l'argomento Attivazione dell'integrazione con IBM InfoSphere Warehouse a pag. 107.)

Opzioni della scheda Server ISW

È possibile specificare la connessione DB2 utilizzata per il caricamento dei dati per la modellazione. Se necessario, inoltre, nella scheda Server è possibile selezionare una connessione specifica per ogni nodo Modelli che sovrascriva la connessione DB2 di default indicata nella finestra di dialogo Applicazioni di supporto. Per ulteriori informazioni, vedere l'argomento Attivazione dell'integrazione con IBM InfoSphere Warehouse a pag. 107.

Figura 5-7
Scheda Server ISW



La connessione utilizzata per la modellazione può corrispondere o meno a quella impiegata nel nodo di input di uno stream. Per esempio, è possibile utilizzare uno stream che accede ai dati di un database DB2, li scarica in IBM® SPSS® Modeler per la pulitura o altre operazioni di modifica e, infine, li carica su un database DB2 differente per la modellazione.

Il nome della sorgente dati ODBC è incorporato in ogni stream di SPSS Modeler. Se uno stream creato su un determinato host viene eseguito su un host differente, il nome della sorgente dati deve essere lo stesso su entrambi gli host. In alternativa, è possibile selezionare una sorgente dati differente nella scheda Server all'interno di ogni nodo Modelli o di input.

Per ricevere feedback durante la creazione di un modello, utilizzare le seguenti opzioni:

- **Attiva feedback.** Selezionare questa opzione per ricevere feedback durante la creazione di un modello (per default, l'opzione è disattivata).
- **Intervallo di feedback (in secondi).** Specificare con quale frequenza SPSS Modeler recupera feedback sullo stato di avanzamento della creazione dei modelli.

Attiva opzioni avanzate di InfoSphere Warehouse Data Mining. Selezionare questa opzione per abilitare il pulsante Opzioni avanzate, che consente di specificare un numero di opzioni avanzate, tra cui un limite di memoria e un SQL personalizzato. Per ulteriori informazioni, vedere l'argomento Opzioni avanzate a pag. 120.

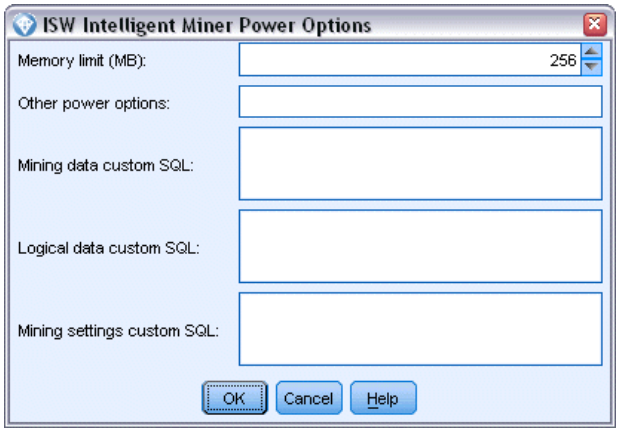
Nella scheda Server di un nodo generato è disponibile un'opzione che consente di eseguire controlli di uniformità grazie all'archiviazione di una stringa identica contenente la chiave del modello generato sia nel modello DB2 che in quello di SPSS Modeler. Per ulteriori informazioni, vedere l'argomento Scheda Server dell'insieme di modelli ISW a pag. 153.

Opzioni avanzate

La scheda Server di tutti gli algoritmi comprende una casella di controllo per l'attivazione delle opzioni avanzate per ISW Modeling. Quando si fa clic sul pulsante Opzioni avanzate, viene visualizzata la finestra di dialogo Opzioni avanzate di ISW, che contiene una serie di opzioni per:

- Limite memoria.
- Altre opzioni avanzate.
- SQL personalizzato data mining.
- SQL personalizzato dati logici.
- SQL personalizzato impostazioni di mining.

Figura 5-8
Impostazioni delle opzioni avanzate di ISW



Limite di memoria. Limita l'utilizzo di memoria di un algoritmo per la creazione di modelli. Si noti che l'opzione avanzata standard imposta un limite sul numero di valori discreti nei dati categoriali.

Altre opzioni avanzate. Consente di definire opzioni avanzate arbitrarie sotto forma di riga di comando per soluzioni o modelli specifici. Le specifiche possono variare in base al tipo di implementazione o soluzione. È possibile estendere manualmente l'SQL generato da IBM® SPSS® Modeler in modo da definire un'operazione di creazione modelli.

SQL personalizzato data mining. È possibile aggiungere chiamate di metodo per modificare l'oggetto DM_MiningData. Per esempio, se si immette il seguente codice SQL, ai dati utilizzati nella creazione di modelli verrà aggiunto un filtro basato su un campo denominato *Partizione*:

```
..DM_setWhereClause('"Partizione" = 1')
```

SQL personalizzato dati logici. È possibile aggiungere chiamate di metodo per modificare l'oggetto DM_LogicalDataSpec. Per esempio, il seguente codice SQL rimuove un campo dall'insieme di campi utilizzato per la creazione di modelli:

```
..DM_remDataSpecFld('campo6')
```

SQL personalizzato impostazioni di mining. È possibile aggiungere chiamate di metodo per modificare l'oggetto DM_ClasSettings/DM_RuleSettings/DM_ClusSettings/DM_RegSettings. Per esempio, se si immette il seguente codice SQL, IBM InfoSphere Warehouse Data Mining

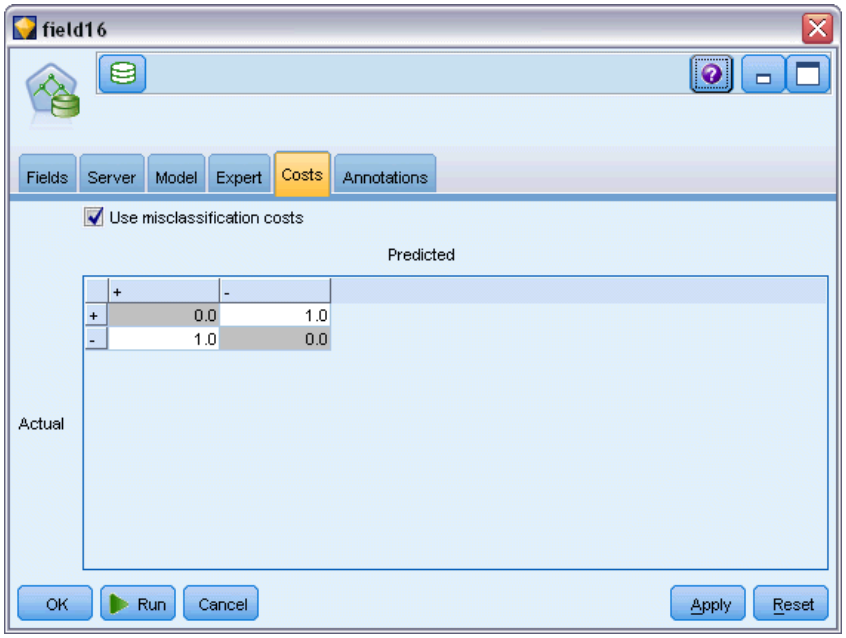
imposterà il campo *Partizione* su attivo (il che significa che verrà sempre incluso nel modello risultante):

```
..DM_setFldUsageType('Partizione',1)
```

Opzioni dei costi di ISW

Nella scheda Costi è possibile modificare i costi di errata classificazione, in maniera da specificare l'importanza relativa dei diversi tipi di errori di previsione.

Figura 5-9
Scheda Costi ISW



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.

Albero decisionale ISW

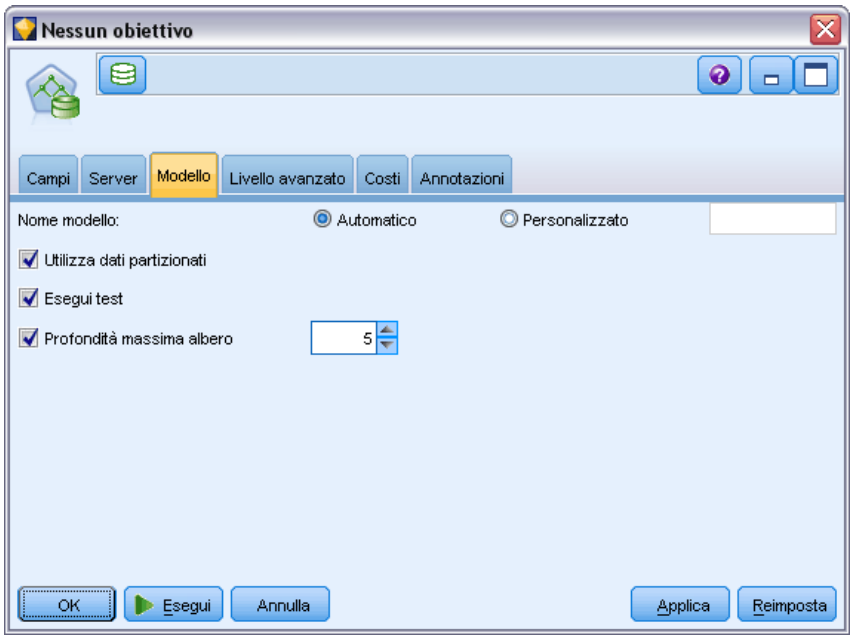
I modelli Alberi decisionali consentono di sviluppare sistemi di classificazione in grado di prevedere o classificare le osservazioni future in base a un insieme di regole decisionali. Se i dati sono divisi in classi di interesse (per esempio, prestiti a basso vs. alto rischio, sottoscrittore vs. non 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 un albero che classifica il rischio sul credito o l'intento di acquisto in base all'età e ad altri fattori.

L'algoritmo per alberi decisionali ISW crea alberi di classificazione su dati di input categoriali. L'albero decisionale risultante è binario. Per la creazione del modello, è possibile applicare diverse impostazioni, compresi i costi di errata classificazione.

Lo strumento ISW Visualizer è l'unico modo per sfogliare i modelli di IBM InfoSphere Warehouse Data Mining.

Opzioni della scheda Modello per il nodo Albero decisionale ISW

Figura 5-10
Scheda Modello del nodo Albero decisionale ISW



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.

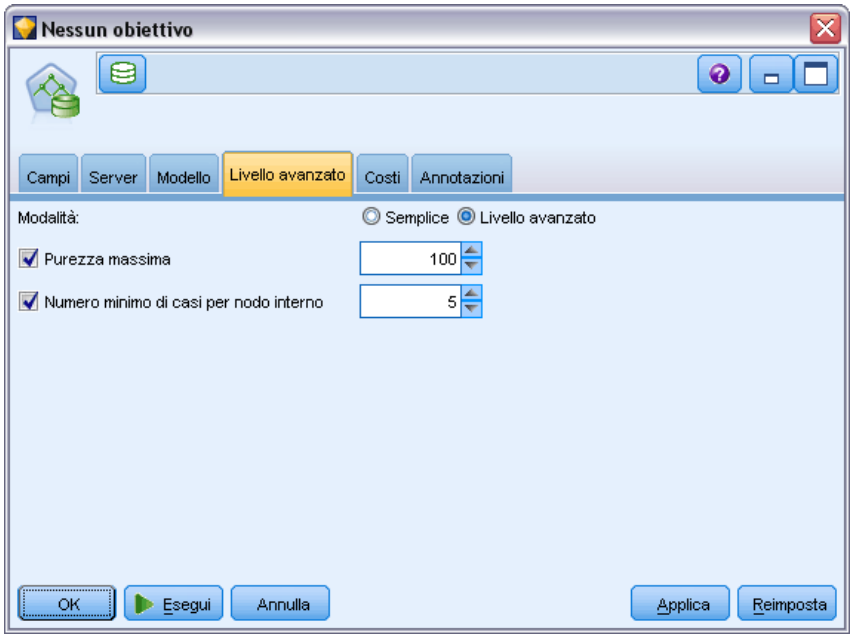
Utilizza dati partizionati. Se si definisce un campo di partizione, selezionare l'opzione Utilizza dati partizionati.

Esegui test. Se si seleziona questa opzione, dopo la creazione del modello nella partizione di addestramento verrà eseguito un test IBM InfoSphere Warehouse Data Mining. Ciò comporterà un passaggio sulla partizione di test per stabilire informazioni su qualità del modello, grafici lift e così via.

Profondità massima albero. È possibile specificare la profondità massima di un albero. In tal modo, si stabilisce che la profondità dell'albero non potrà superare il numero specificato di livelli. Se questa opzione non viene selezionata, non verrà applicato alcun limite. Per evitare modelli eccessivamente complessi, si sconsiglia, se non in caso di stretta necessità, di definire un valore superiore a 5.

Opzioni avanzate Albero decisionale ISW

Figura 5-11
Scheda Livello avanzato del nodo Albero decisionale ISW



Purezza massima. Questa opzione imposta la purezza massima per i nodi interni. Qualora, in seguito alla suddivisione di un nodo, uno dei figli superi la misura di purezza definita (se, per esempio, oltre il 90% dei casi rientra in una categoria specificata), il nodo non verrà suddiviso.

Numero minimo di casi per nodo interno. Se il processo di suddivisione comporta la creazione di un nodo con un numero di casi inferiore al minimo specificato, il nodo non verrà suddiviso.

Associazione ISW

Il nodo Associazione ISW può essere utilizzato per trovare delle regole di associazione tra gli elementi presenti in un insieme di gruppi. Le regole di associazione consentono di associare una conclusione specifica, per esempio l'acquisto di un particolare prodotto, a un insieme di condizioni, come l'acquisto di numerosi altri prodotti.

È possibile decidere di includere o escludere le regole di associazione dal modello specificando dei **vincoli**. Se si decide di includere un determinato campo di input, vengono incluse nel modello le regole di associazione contenenti almeno uno degli elementi specificati. Se si esclude un campo di input, le regole di associazione che contengono uno qualsiasi degli elementi specificati vengono eliminate dai risultati.

Gli algoritmi di associazione e sequenza di ISW prevedono l'utilizzo di **tassonomie**. Queste ultime mappano singoli valori a concetti di livello superiore. Per esempio, penne e matite possono essere mappate a una categoria cancelleria.

Le regole di associazione hanno un solo conseguente (la conclusione) e più antecedenti (l'insieme di condizioni). Esempio:

[Pane, Marmellata] • [Burro]

[Pane, Marmellata]
• [Margarina]

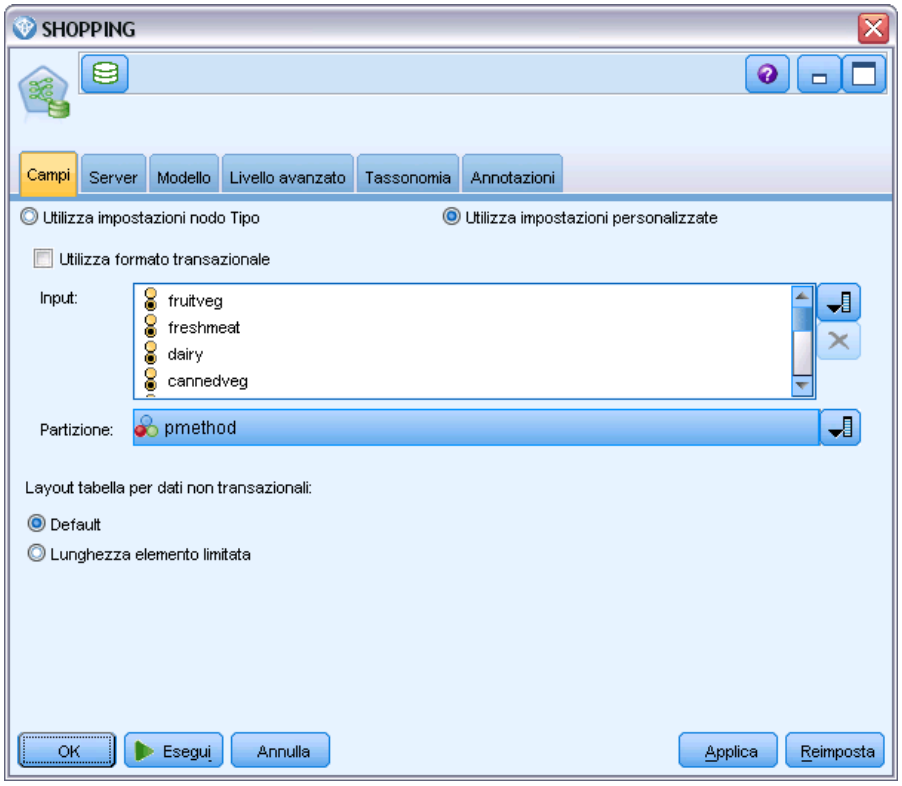
Qui, Bread e Jam sono gli antecedenti (detti anche **corpo della regola**) e Butter o Margarine sono altrettanti esempi di conseguenti (detti anche **intestazione della regola**). La prima regola indica che una persona che ha acquistato pane e marmellata ha acquistato contemporaneamente anche del burro. La seconda regola identifica un cliente che al momento dell'acquisto della stessa combinazione (pane e marmellata) ha acquistato anche margarina nello stesso negozio.

Lo strumento visualizzatore è l'unico modo per sfogliare i modelli di IBM InfoSphere Warehouse Data Mining.

Opzioni dei campi Associazione ISW

Nella scheda Campi si indicano i campi da utilizzare nella creazione del modello.

Figura 5-12
Scheda Campi del nodo Associazione ISW



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. Oltre all'impostazione di default relativa all'utilizzo del nodo Tipo per la selezione dei campi obiettivo e di input, in questa scheda è possibile modificare soltanto l'opzione relativa al layout di tabella per i dati non transazionali.

Utilizza impostazioni nodo Tipo. Questa opzione specifica l'utilizzo delle informazioni sui campi da un nodo Tipo a monte. È l'impostazione di default.

Utilizza impostazioni personalizzate. Questa opzione specifica l'utilizzo delle informazioni sui campi immessi 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.

Utilizzo del formato transazionale. Selezionare la 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, vedere l'argomento Dati tabulari e dati transazionali in il capitolo 12 in *IBM SPSS Modeler 15 Nodi Modelli*.

- **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).
- **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 un unico campo nominale quando i dati sono in formato transazionale.

Utilizzo del formato tabulare. Deselezionare la casella di controllo Utilizza formato transazionale se i dati di origine sono in formato tabulare.

- **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 validazione 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 di partizione utilizzando nodi Tipo o Partizione, nella scheda Campi di ogni nodo Modelli che utilizza il partizionamento è necessario selezionare un singolo campo di partizione. Se è presente un'unica partizione, verrà utilizzata automaticamente quando si attiva il partizionamento. Per ulteriori informazioni, vedere l'argomento Nodo Partizione in il capitolo 4 in *IBM SPSS Modeler 15 Nodi di input, elaborazione e output*. 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.

Layout tabella per dati non transazionali. Per i dati tabulari, è possibile scegliere un layout di tabella standard (opzione di default) o layout a lunghezza degli elementi limitata.

Nel layout di default, il numero di colonne è determinato dal numero totale di elementi associati.

Tabella 5-2
Layout di tabella di default

ID gruppo	Conto corrente	Conto di risparmio	Carta di credito	Prestito	Deposito titoli
Smith	Y	Y	Y	-	-
Jackson	Y	-	Y	Y	Y
Douglas	Y	-	-	-	Y

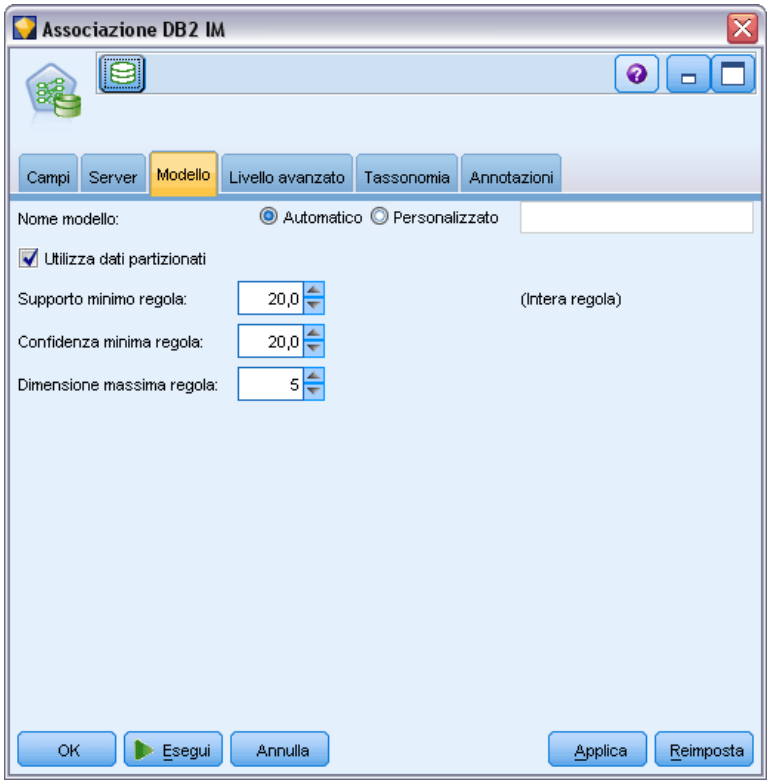
Nel layout a lunghezza degli elementi limitata, il numero di colonne è determinato dal numero più alto di elementi associati in una qualsiasi delle righe.

Tabella 5-3
Layout di tabella a lunghezza degli elementi limitata

ID gruppo	Elemento1	Elemento2	Elemento3	Elemento4
Smith	conto corrente	conto di risparmio	carta di credito	-
Jackson	conto corrente	carta di credito	prestito	deposito titoli
Douglas	conto corrente	deposito titoli	-	-

Opzioni della scheda Modello per il nodo Associazione ISW

Figura 5-13
Scheda Modello del nodo Associazione ISW



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 di partizione, questa opzione garantisce che per la creazione del modello verranno utilizzati solo i dati della partizione di addestramento. Per ulteriori informazioni, vedere l'argomento Nodo Partizione in il capitolo 4 in *IBM SPSS Modeler 15 Nodi di input, elaborazione e output*.

Supporto minimo regola (%). Livello di supporto minimo per le regole di associazione o di sequenza. Vengono incluse nel modello solo le regole che raggiungono almeno questo livello di supporto. Il valore è calcolato come A/B*100, dove A è il numero dei gruppi che contengono tutti gli elementi presenti nella regola e B è il numero totale dei gruppi considerati. Per prendere in esame le associazioni o le sequenze più comuni, aumentare il valore di questa impostazione.

Confidenza minima regola (%). Livello di confidenza minimo per le regole di associazione o di sequenza. Vengono incluse nel modello solo le regole che raggiungono almeno questo livello di confidenza. Il valore è calcolato come *m*/*n*100, dove *m* è il numero dei gruppi che contengono l'intestazione della regola (conseguente) unita al corpo della regola (antecedente), e *n* è il numero dei gruppi che contengono il corpo della regola. Se si ottengono associazioni o sequenze in numero eccessivo o non interessanti, provare ad aumentare il valore di questa impostazione.

Se le associazioni o le sequenze ottenute sono troppo poche, provare a diminuire il valore di questa impostazione.

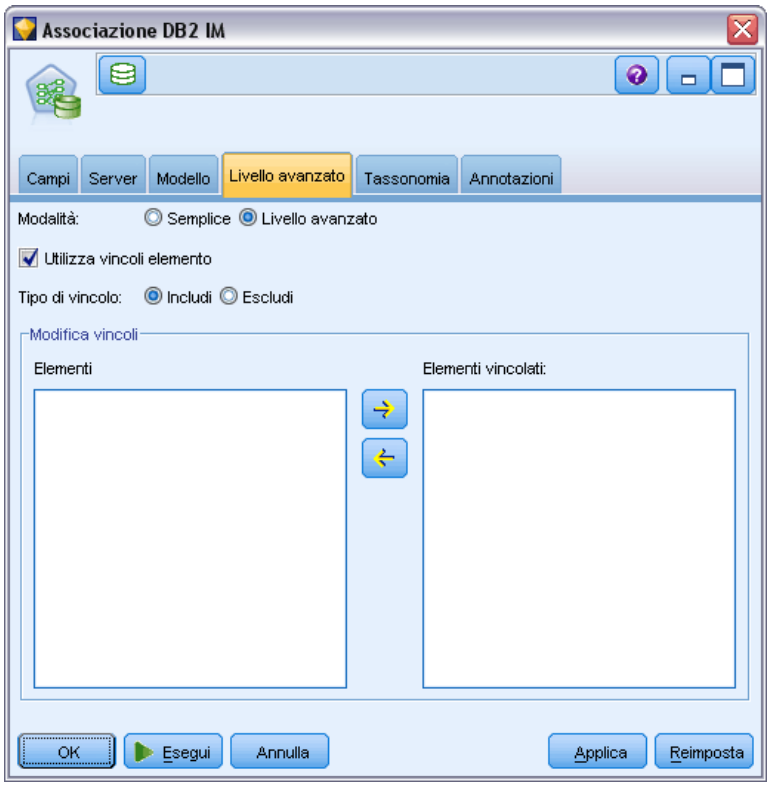
Dimensione massima regola. Numero massimo di elementi consentito in una regola, compreso l'elemento conseguente. Se le associazioni o le sequenze interessanti sono relativamente brevi, è possibile diminuire il valore di questa impostazione per accelerare la creazione dell'insieme.

Nota: il punteggio viene calcolato solo per i nodi con dati di input transazionali; le tabelle di verità (dati tabulari) rimangono grezze.

Opzioni della scheda Opzioni avanzate per il nodo Associazione ISW

Sulla scheda Livello avanzato del nodo Associazione è possibile indicare le regole di associazione da includere o da escludere dai risultati. Se si decide di includere elementi specificati, vengono incluse nel modello le regole contenenti almeno uno degli elementi specificati. Se si decide di escludere gli elementi specificati, le regole che contengono uno qualsiasi degli elementi specificati vengono eliminate dai risultati.

Figura 5-14
Scheda Livello avanzato del nodo Associazione ISW



Se si seleziona Utilizza vincoli elemento, tutti gli elementi aggiunti all'elenco dei vincoli verranno inclusi o esclusi dai risultati, a seconda dell'impostazione specificata per Tipo di vincolo.

Tipo di vincolo. Decidere se includere o escludere dai risultati le regole di associazione che contengono gli elementi specificati.

Modifica vincoli. Per aggiungere un elemento all'elenco degli elementi vincolati, selezionarlo nell'elenco Elementi e fare clic sul pulsante freccia destra.

Opzioni della scheda Tassonomia per ISW

Gli algoritmi di associazione e sequenza di ISW prevedono l'utilizzo di **tassonomie**. Queste ultime mappano singoli valori a concetti di livello superiore. Per esempio, penne e matite possono essere mappate a una categoria cancelleria.

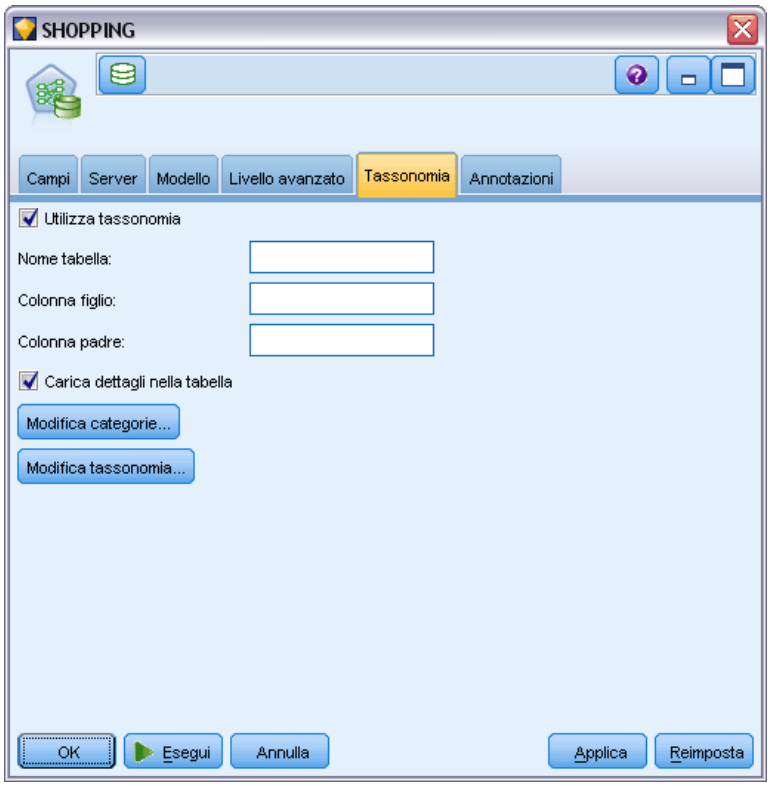
Nella scheda Tassonomia è possibile definire mappe di categorie per esprimere tassonomie sui dati. Per esempio, una tassonomia può creare due categorie (Staple e Luxury) e quindi assegnare gli elementi fondamentali a ognuna di esse. Per esempio, wine è assegnato a Luxury e bread è assegnato a Staple. La tassonomia ha una struttura padre-figlio, come la seguente:

Figlio	Genitore
vino	Prodotti di lusso
pane	Prodotti alimentari di base

Utilizzando questa tassonomia, è possibile creare un modello di associazione o di sequenza comprendente regole che coinvolgono sia le categorie che gli elementi fondamentali.

Nota: per attivare le opzioni di questa scheda è necessario che i dati di input siano in formato transazionale e occorre selezionare Utilizza formato transazionale nella scheda Campi e quindi Utilizza tassonomia in questa scheda. Per ulteriori informazioni, vedere l'argomento Dati tabulari e dati transazionali in il capitolo 12 in *IBM SPSS Modeler 15 Nodi Modelli*.

Figura 5-15
Scheda Tassonomia del nodo Associazione ISW



Nome tabella. Consente di specificare il nome della tabella DB2 in cui archiviare i dettagli relativi alla tassonomia.

Colonna figlio. Consente di specificare il nome della colonna figlio nella tabella di tassonomia. Tale colonna contiene i nomi di elemento o i nomi di categoria.

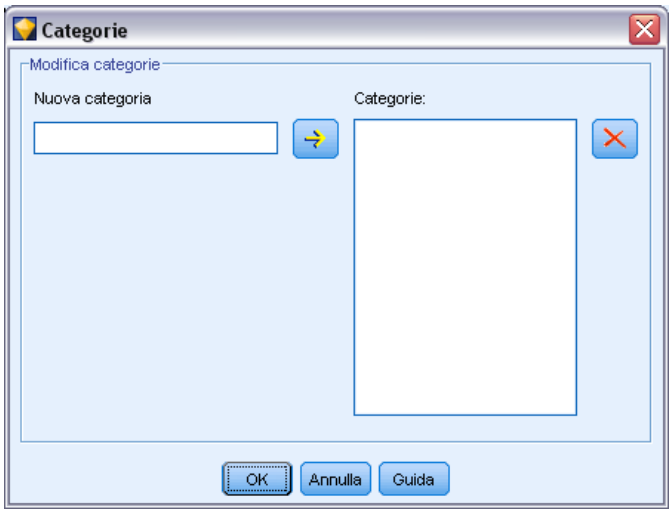
Colonna padre. Consente di specificare il nome della colonna padre nella tabella di tassonomia. Tale colonna contiene i nomi di categoria.

Carica dettagli nella tabella. Selezionare questa opzione se le informazioni sulla tassonomia archiviate in IBM® SPSS® Modeler devono essere caricate nella tabella di tassonomia alla creazione del modello. Si noti che se già esiste una tabella di tassonomia, tale tabella verrà eliminata. Le informazioni di tassonomia vengono memorizzate con il nodo di creazione del modello e possono essere modificate utilizzando i pulsanti Modifica categorie e Modifica tassonomia.

Editor di categorie

La finestra di dialogo Modifica categorie consente di aggiungere o di eliminare categorie da un elenco ordinato.

Figura 5-16
Editor di categorie tassonomia



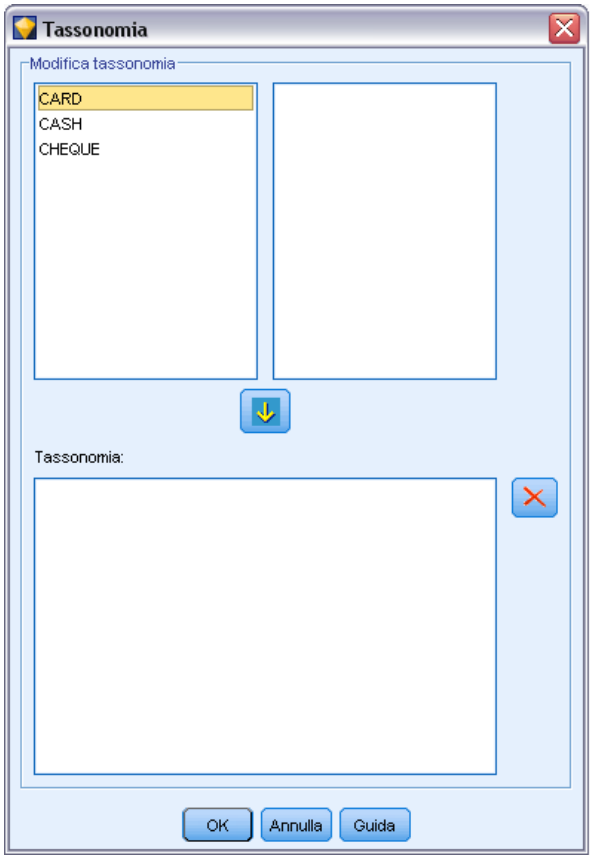
Per aggiungere una categoria, digitarne il nome nel campo Nuova categoria e fare clic sul pulsante freccia per spostarla nell'elenco Categorie.

Per rimuovere una categoria, selezionarla nell'elenco Categorie e fare clic sul pulsante Elimina adiacente.

Editor di tassonomia

La finestra di dialogo Modifica tassonomia consente di combinare l'insieme di elementi fondamentali definiti nei dati e l'insieme di categorie per creare una tassonomia. Per aggiungere nuove voci alla tassonomia, selezionare uno o più elementi o una o più categorie dall'elenco a sinistra e una o più categorie dall'elenco a destra, quindi fare clic sul pulsante con la freccia. Si noti che, qualora eventuali aggiunte alla tassonomia producano un conflitto (per esempio, se si specifica sia `cat1 -> cat2` che l'opposto, `cat2 -> cat1`), tali aggiunte non verranno effettuate.

Figura 5-17
Editor di tassonomia



Sequenza ISW

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 **insiemi 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 insiemi di elementi. Il primo insieme di elementi contiene il pane e il latte e il secondo contiene il formaggio. Per **sequenza** si intende un elenco di insiemi 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.

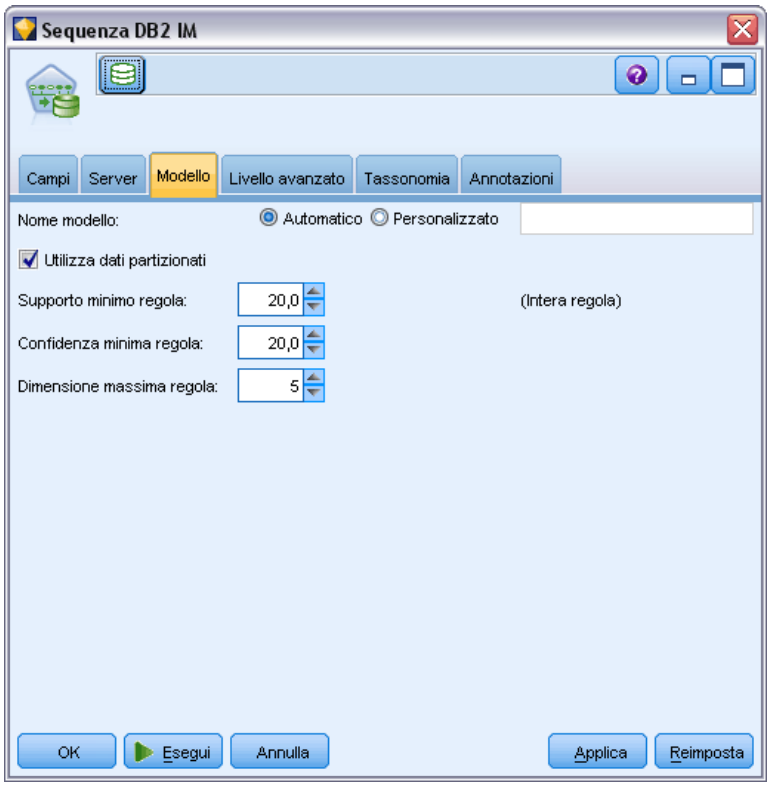
È possibile usare la funzione mining Regole di sequenza in diverse aree di attività. Per esempio, nel settore delle vendite al dettaglio è possibile trovare insiemi di acquisti tipici. Questi insiemi mostrano le diverse combinazioni di clienti, prodotti e ora dell'acquisto. Mediante queste informazioni, è possibile identificare i clienti potenziali di un prodotto che non hanno ancora acquistato il prodotto. Inoltre, è possibile offrire prodotti ai clienti potenziali nei tempi previsti.

Una sequenza rappresenta una serie ordinata di insiemi di elementi. Le sequenze contengono i seguenti livelli di raggruppamento:

- Gli eventi che accadono simultaneamente formano un'unica transazione o un insieme di elementi.
- Ogni elemento o insieme di elementi appartiene a un gruppo di transazioni. Per esempio, un articolo acquistato appartiene a un cliente, un clic su una pagina specifica appartiene a un utente Web, un componente appartiene a un'automobile prodotta. Diversi insiemi di elementi che avvengono in momenti diversi e appartengono allo stesso gruppo di transazioni formano una sequenza.

Opzioni della scheda Modello per il nodo Sequenza ISW

Figura 5-18
Scheda Modello del nodo Sequenza ISW



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 di partizione, questa opzione garantisce che per la creazione del modello verranno utilizzati solo i dati della partizione di addestramento. Per ulteriori informazioni, vedere l'argomento Nodo Partizione in il capitolo 4 in *IBM SPSS Modeler 15 Nodi di input, elaborazione e output*.

Supporto minimo regola (%). Livello di supporto minimo per le regole di associazione o di sequenza. Vengono incluse nel modello solo le regole che raggiungono almeno questo livello di supporto. Il valore è calcolato come A/B*100, dove A è il numero dei gruppi che contengono tutti gli elementi presenti nella regola e B è il numero totale dei gruppi considerati. Per prendere in esame le associazioni o le sequenze più comuni, aumentare il valore di questa impostazione.

Confidenza minima regola (%). Livello di confidenza minimo per le regole di associazione o di sequenza. Vengono incluse nel modello solo le regole che raggiungono almeno questo livello di confidenza. Il valore è calcolato come *m*/*n**100, dove *m* è il numero dei gruppi che contengono l'intestazione della regola (conseguente) unita al corpo della regola (antecedente), e *n* è il numero dei gruppi che contengono il corpo della regola. Se si ottengono associazioni o sequenze in numero eccessivo o non interessanti, provare ad aumentare il valore di questa impostazione. Se le associazioni o le sequenze ottenute sono troppo poche, provare a diminuire il valore di questa impostazione.

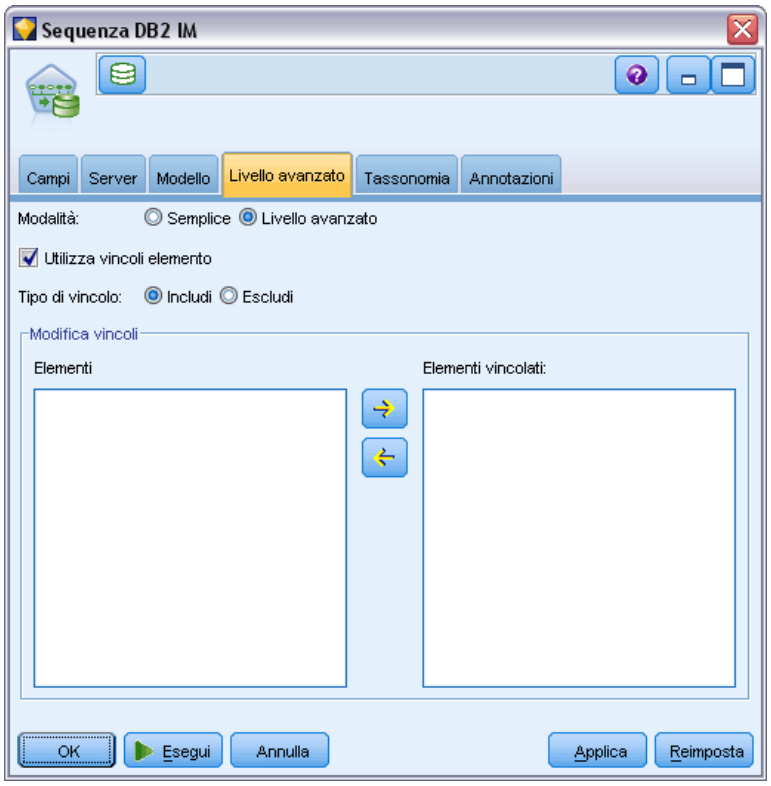
Dimensione massima regola. Numero massimo di elementi consentito in una regola, compreso l'elemento conseguente. Se le associazioni o le sequenze interessanti sono relativamente brevi, è possibile diminuire il valore di questa impostazione per accelerare la creazione dell'insieme.

Nota: il punteggio viene calcolato solo per i nodi con dati di input transazionali; le tabelle di verità (dati tabulari) rimangono grezze.

Opzioni della scheda Opzioni avanzate per il nodo Sequenza ISW

È possibile specificare le regole di sequenza da includere o da escludere dai risultati. Se si decide di includere elementi specificati, vengono incluse nel modello le regole contenenti almeno uno degli elementi specificati. Se si decide di escludere gli elementi specificati, le regole che contengono uno qualsiasi degli elementi specificati vengono eliminate dai risultati.

Figura 5-19
Scheda Livello avanzato del nodo Sequenza ISW



Se si seleziona Utilizza vincoli elemento, tutti gli elementi aggiunti all'elenco dei vincoli verranno inclusi o esclusi dai risultati, a seconda dell'impostazione specificata per Tipo di vincolo.

Tipo di vincolo. Decidere se includere o escludere dai risultati le regole di associazione che contengono gli elementi specificati.

Modifica vincoli. Per aggiungere un elemento all'elenco degli elementi vincolati, selezionarlo nell'elenco Elementi e fare clic sul pulsante freccia destra.

Regressione ISW

Il nodo Regressione ISW supporta i seguenti algoritmi di regressione:

- Trasformazioni (default).
- Lineare
- Polynomial
- RBF

Regressione trasformazione

L'algoritmo di regressione trasformazione ISW consente di creare modelli che rappresentano alberi decisionali con equazioni di regressione in corrispondenza delle tre foglie. Si noti che la struttura di tali modelli non potrà essere visualizzata mediante lo strumento Visualizer di IBM.

Il browser di IBM® SPSS® Modeler visualizza le impostazioni e le annotazioni, ma non la struttura dei modelli. L'algoritmo prevede un numero relativamente esiguo di impostazioni configurabili dall'utente.

Regressione lineare

L'algoritmo di regressione lineare ISW presuppone che esista una relazione lineare tra i campi esplicativi e il campo obiettivo e genera modelli che rappresentano delle equazioni. È normale che il valore previsto sia diverso dal valore osservato, in quanto un'equazione di regressione rappresenta un'approssimazione del campo obiettivo. La differenza viene chiamata residuo.

Il modello di IBM InfoSphere Warehouse Data Mining riconosce i campi che non presentano un valore esplicativo. Per stabilire se un campo dispone di un valore esplicativo, l'algoritmo di regressione lineare esegue dei test statistici oltre alla selezione di variabili autonomiche. Se si conoscono già i campi che non presentano un valore esplicativo, è possibile selezionare automaticamente un sottoinsieme dei campi esplicativi in modo da ridurre i tempi di esecuzione.

L'algoritmo di regressione lineare offre i seguenti metodi per selezionare automaticamente i sottoinsiemi di campi esplicativi:

Regressione stepwise. Per utilizzare il metodo di regressione stepwise, è necessario specificare un livello di significatività minima. Solo i campi che presentano un livello di significatività superiore al valore specificato vengono utilizzati dall'algoritmo di regressione lineare.

Regressione di R-quadrato. Il metodo di regressione di R-quadrato identifica il modello ideale mediante l'ottimizzazione di una misura di qualità del modello. Viene utilizzata una delle seguenti misure di qualità:

- Coefficiente di correlazione di Pearson quadrato
- Coefficiente di correlazione di Pearson quadrato corretto.

Per default, l'algoritmo di regressione lineare seleziona automaticamente i sottoinsiemi dei campi esplicativi, utilizzando il coefficiente di correlazione di Pearson quadrato corretto per ottimizzare la qualità del modello.

Regressione polinomiale

L'algoritmo di regressione polinomiale ISW presuppone una relazione polinomiale. Un modello di regressione polinomiale è un'equazione formata dalle seguenti parti:

- Il massimo grado di regressione polinomiale
- Un'approssimazione del campo obiettivo
- I campi esplicativi.

Regressione RBF

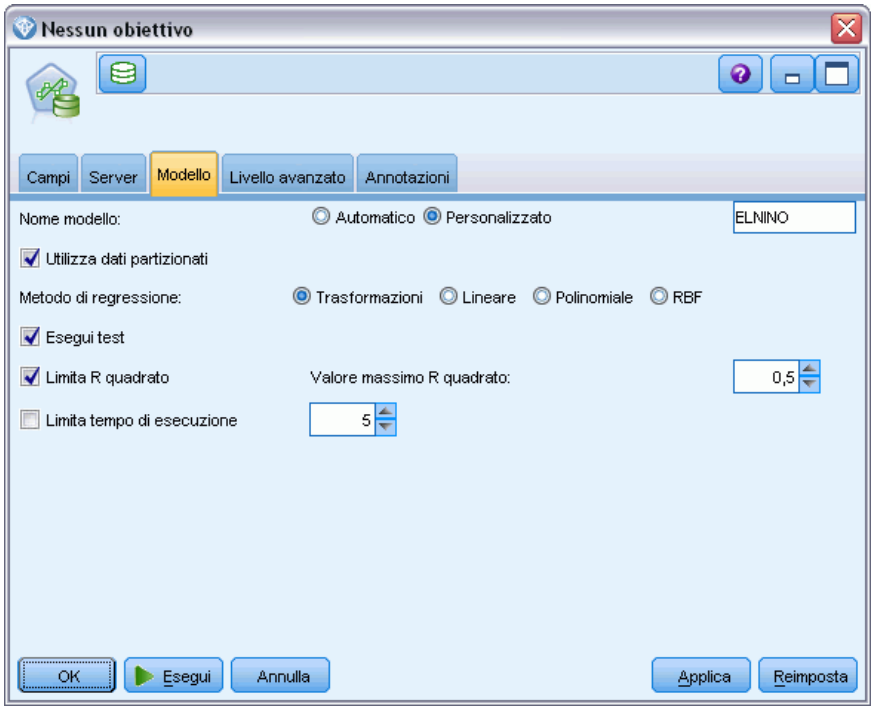
L'algoritmo di regressione RBF ISW presuppone che esista una relazione tra i campi esplicativi e il campo obiettivo. Questa relazione si può esprimere come una combinazione lineare di funzioni gaussiane. Le funzioni gaussiane sono funzioni RBF specifiche.

Opzioni della scheda Modello per il nodo Regressione ISW

Nella scheda Modello del nodo Regressione ISW è possibile specificare il tipo di algoritmo di regressione da utilizzare, oltre a:

- Se utilizzare o meno dati partizionati
- Se eseguire o meno un test
- Un limite per il valore R^2
- Un limite per il tempo di esecuzione

Figura 5-20
Scheda Modello del nodo Regressione ISW



Utilizza dati partizionati. Se è definito un campo di partizione, questa opzione garantisce che per la creazione del modello verranno utilizzati solo i dati della partizione di addestramento. Per ulteriori informazioni, vedere l'argomento Nodo Partizione in il capitolo 4 in *IBM SPSS Modeler 15 Nodi di input, elaborazione e output*.

Metodo di regressione. Scegliere il tipo di regressione da eseguire. Per ulteriori informazioni, vedere l'argomento Regressione ISW a pag. 136.

Esegui test. Se si seleziona questa opzione, dopo la creazione del modello nella partizione di addestramento verrà eseguito un test InfoSphere Warehouse Data Mining. Ciò comporterà un passaggio sulla partizione di test per stabilire informazioni su qualità del modello, grafici lift e così via.

Limita R quadrato. Questa opzione specifica l'errore sistematico massimo tollerato (il coefficiente di correlazione di Pearson quadrato, R^2). Questo coefficiente misura la correlazione tra l'errore di previsione sui dati di verifica e i valori obiettivo effettivi. Ha un valore compreso tra 0 (nessuna correlazione) e 1 (correlazione positiva o negativa perfetta). Il valore definito qui imposta il limite massimo per l'errore sistematico tollerato del modello.

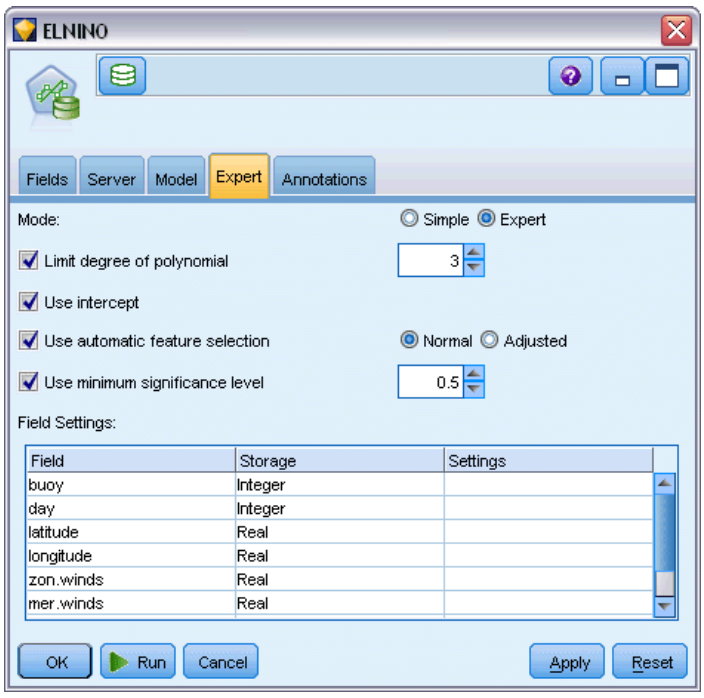
Limita tempo di esecuzione. Specificare in minuti il tempo di esecuzione massimo desiderato.

Opzioni avanzate del nodo Regressione ISW

Nella scheda Livello avanzato del nodo Regressione ISW è possibile specificare una serie di opzioni avanzate per la regressione lineare, polinomiale o RBF.

Opzioni avanzate per la regressione lineare o polinomiale

Figura 5-21
Scheda Livello avanzato del nodo Regressione ISW per la regressione lineare o polinomiale



Limita grado di polinomiale. Imposta il limite massimo di regressione polinomiale. Se si imposta il grado di regressione polinomiale massimo su 1, l'algoritmo di regressione polinomiale è identico a quello della regressione lineare. Se si specifica un valore alto per il grado massimo di regressione polinomiale, l'algoritmo di regressione polinomiale tenderà a sovradattare i dati. Questo significa

che il modello risultante approssimerà in modo accurato i dati di addestramento, ma non risulterà valido se applicato a dati non utilizzati per l'addestramento.

Utilizza intercetta. Quando è attivata questa opzione, la curva di regressione viene forzata a passare attraverso l'origine. Questo significa che il modello non conterrà un termine costante.

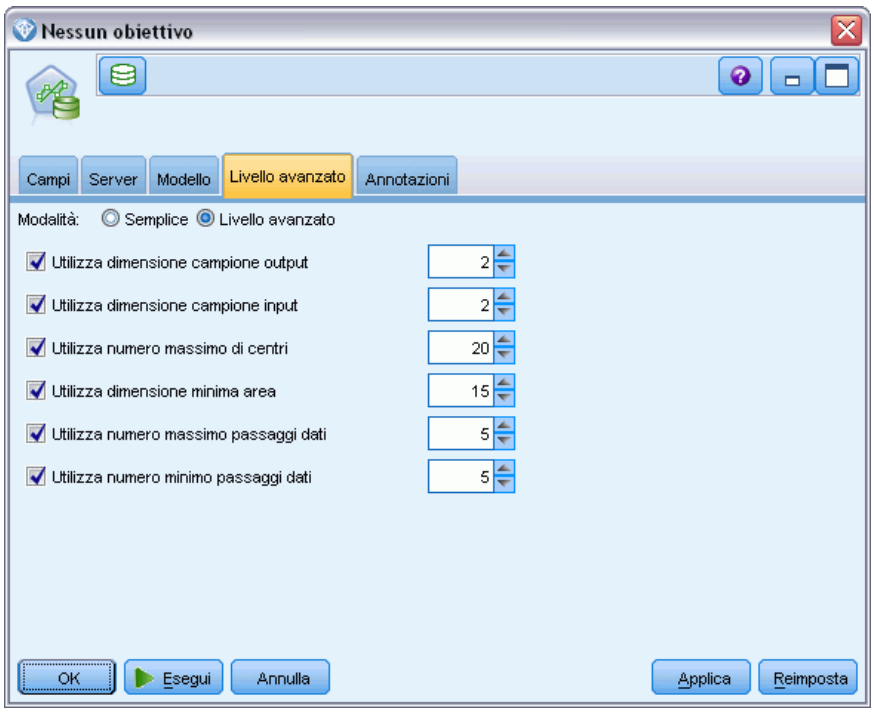
Utilizza selezione automatica delle funzionalità. Quando è attivata questa opzione, l'algoritmo tenta di determinare un sottoinsieme ottimale di possibili predittori se non si specifica un livello minimo di significatività.

Utilizza livello minimo di significatività. Quando si specifica un livello minimo di significatività, viene utilizzata la regressione stepwise per determinare un sottoinsieme di possibili predittori. Solo i campi indipendenti la cui significatività è superiore al valore specificato contribuiscono al calcolo del modello di regressione.

Impostazioni campo. Per specificare le opzioni per i singoli campi di input, fare clic sulla riga corrispondente nella colonna Impostazioni della tabella Impostazioni campo e scegliere <Specifica impostazioni>. Per ulteriori informazioni, vedere l'argomento Specifica delle impostazioni dei campi per la regressione a pag. 141.

Opzioni avanzate per la regressione RBF

Figura 5-22
Scheda Livello avanzato del nodo Regressione ISW per la regressione RBF



Utilizza dimensione campione output. Definisce un campione 1-ogni-N per la verifica e il test del modello.

Utilizza dimensione campione input. Definisce un campione 1-ogni-N per l'addestramento.

Utilizza numero massimo di centri. Il numero massimo di centri creati a ogni passaggio. Dal momento che il numero dei centri può aumentare fino al doppio del numero iniziale durante un passaggio, il numero effettivo di centri può essere superiore al numero specificato.

Utilizza dimensione minima area. Il numero minimo di record assegnati a un'area.

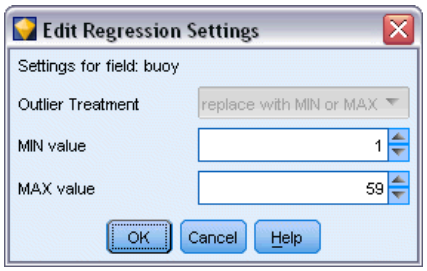
Utilizza numero massimo passaggi dati. Il numero massimo di passaggi effettuato dall'algoritmo nei dati di input. Se specificato, questo valore deve essere maggiore o uguale al numero minimo di passaggi.

Utilizza numero minimo passaggi dati. Il numero minimo di passaggi effettuato dall'algoritmo nei dati di input. Specificare un valore elevato solo se si dispone di dati di addestramento sufficienti e se si è certi dell'esistenza di un buon modello.

Specifica delle impostazioni dei campi per la regressione

Qui è possibile specificare l'intervallo di valori per un singolo campo di input.

Figura 5-23
Specifica delle impostazioni di regressione per un campo di input



Valore MIN. Valore valido minimo del campo di input.

Valore MAX. Valore valido massimo del campo di input.

Raggruppamento cluster ISW

La funzione di mining di raggruppamento tramite cluster cerca nei dati di input le caratteristiche comuni che ricorrono più frequentemente e raggruppa i dati di input in cluster. I membri di ciascun cluster hanno proprietà simili. Non vengono applicate nozioni preconcette su quali schemi esistano all'interno dei dati. Il raggruppamento tramite cluster è un processo di individuazione.

Il nodo Raggruppamento cluster ISW offre i seguenti metodi di raggruppamento tramite cluster:

- Demografico
- Kohonen
- BIRCH ottimizzato (Balanced Iterative Reducing and Clustering using Hierarchies)

La tecnica dell'algoritmo di **raggruppamento cluster demografici** è basata sulla distribuzione. Il raggruppamento tramite cluster basato sulla distribuzione consente un rapido e semplice raggruppamento dei database molto estesi. Il numero di cluster viene scelto automaticamente

(l'utente può specificare il numero massimo). È disponibile un'ampia gamma di parametri configurabili dall'utente.

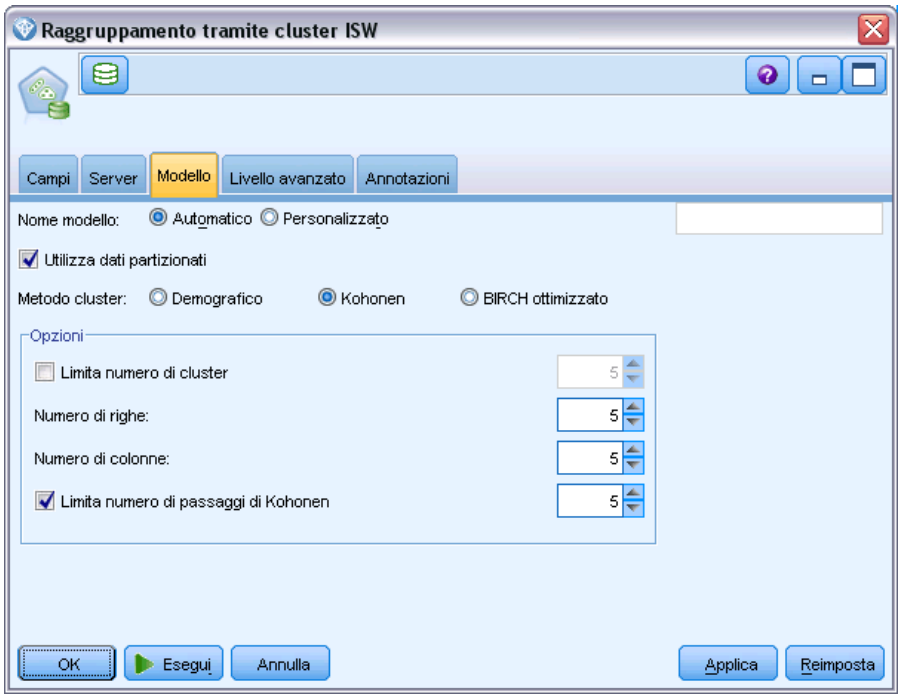
La tecnica dell'algoritmo di **raggruppamento cluster Kohonen** è basata sul centro. La mappa della funzione Kohonen cerca di posizionare i centri dei cluster in modo da ridurre al minimo la distanza totale tra i record e il centro dei cluster. La possibilità di separare i cluster non viene presa in considerazione. I vettori centrali vengono sistemati in una mappa con un determinato numero di colonne e righe. Questi vettori sono interconnessi in modo che oltre al vettore vincente, che si trova più vicino a un record di addestramento, vengano regolati anche tutti i vettori che si trovano nelle vicinanze. Tuttavia, la quantità di regolazione apportata diminuisce con l'aumentare della distanza dei centri.

La tecnica dell'algoritmo ottimizzato del **raggruppamento Birch** è basata sulla distribuzione e mira a ridurre al minimo la distanza totale tra i record e i relativi cluster. La distanza di log-verosimiglianza viene utilizzata per default per determinare la distanza tra un record e un cluster; in alternativa, è possibile selezionare la distanza euclidea se tutti i campi attivi sono numerici. L'algoritmo BIRCH esegue due passaggi distinti: innanzitutto dispone i record di input in un albero CF (Clustering Feature) in modo che i record simili appartengano agli stessi nodi dell'albero, quindi raggruppa le foglie dell'albero in memoria per generare il risultato di raggruppamento finale.

Opzioni della scheda Modello per il nodo Raggruppamento cluster ISW

Nella scheda Modello del nodo Raggruppamento cluster è possibile specificare il metodo da utilizzare per creare i cluster e alcune opzioni correlate.

Figura 5-24
Scheda Modello del nodo Raggruppamento cluster ISW



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 di partizione, questa opzione garantisce che per la creazione del modello verranno utilizzati solo i dati della partizione di addestramento. Per ulteriori informazioni, vedere l'argomento Nodo Partizione in il capitolo 4 in *IBM SPSS Modeler 15 Nodi di input, elaborazione e output*.

Metodo cluster. Scegliere il metodo da utilizzare per creare i cluster: Demografico, Kohonen o BIRCH ottimizzato. Per ulteriori informazioni, vedere l'argomento Raggruppamento cluster ISW a pag. 141.

Limita numero di cluster. La limitazione del numero di cluster consente di risparmiare tempo in fase di esecuzione evitando la produzione di molti cluster di piccole dimensioni.

Numero di righe/Numero di colonne. (Solo per il metodo Kohonen) Specifica il numero di righe e di colonne della mappa delle caratteristiche Kohonen. (Disponibile solo se è selezionato Limita numero di passaggi di Kohonen ed è deselezionato Limita numero di cluster.)

Limita numero di passaggi di Kohonen. (Solo per il metodo Kohonen) Specifica il numero di passaggi compiuti dall'algoritmo di raggruppamento tramite cluster sui dati durante le esecuzioni dell'addestramento. A ogni passaggio, i vettori del centro vengono regolati in modo da ridurre al minimo la distanza tra i centri dei cluster e i record. Inoltre, a ogni passaggio, diminuisce la quantità di regolazione apportata ai vettori. Al primo passaggio, le regolazioni sono approssimative. All'ultimo passaggio, l'entità della regolazione apportata ai centri è piuttosto ridotta. Vengono eseguite solo regolazioni minime.

Misura della distanza. (Solo per il metodo BIRCH ottimizzato) Selezionare la misura della distanza dal record al cluster utilizzata dall'algoritmo BIRCH. È possibile selezionare la distanza di verosimiglianza, che rappresenta l'importazione di default, oppure la distanza euclidea. *Nota*: la distanza euclidea può essere selezionata soltanto se tutti i campi attivi sono numerici.

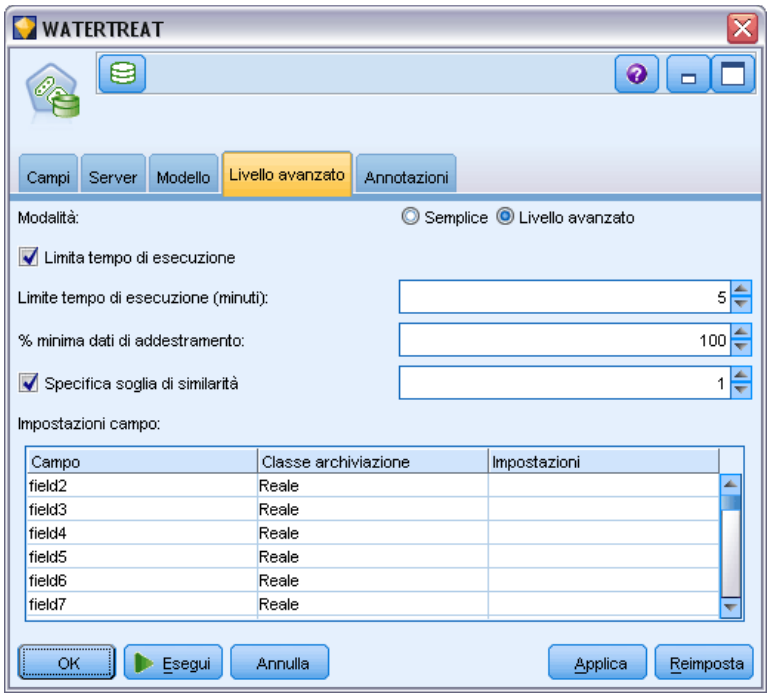
Numero massimo di nodi foglia. (Solo per il metodo BIRCH ottimizzato) Numero massimo di nodi foglia che si desidera includere nell'albero CF (Clustering Feature). L'albero CF (Clustering Feature) è il risultato del primo passaggio dell'algoritmo BIRCH ottimizzato, nel quale i record di dati sono disposti in un albero in modo che i record simili appartengano allo stesso nodo foglia. Il tempo di esecuzione dell'algoritmo aumenta con il numero dei nodi foglia. Il valore di default è 1000.

Passaggi Birch. (Solo per il metodo BIRCH ottimizzato) Numero di passaggi eseguiti dall'algoritmo sui dati per perfezionare il risultato del raggruppamento in cluster. Il numero di passaggi ha effetto sul tempo di elaborazione delle esecuzioni di addestramento (poiché ciascun passaggio richiede una scansione completa dei dati) e sulla qualità del modello. Valori bassi limitano il tempo di elaborazione, ma possono ridurre la qualità dei modelli. Valori alti aumentano il tempo di elaborazione ma generano solitamente modelli migliori. In media, 3 o più passaggi producono buoni risultati. Il valore predefinito è 3.

Opzioni avanzate del nodo Raggruppamento cluster ISW

Nella scheda Livello avanzato del nodo Raggruppamento cluster è possibile specificare opzioni avanzate come soglie di similarità, limiti per il tempo di esecuzione e pesi dei campi.

Figura 5-25
Scheda Livello avanzato del nodo Raggruppamento cluster ISW



Limita tempo di esecuzione. Selezionare questa casella per abilitare le opzioni che consentono di controllare il tempo impiegato per creare il modello. È possibile specificare un intervallo di tempo in minuti, una percentuale minima di dati di addestramento da elaborare o entrambi i valori. Per il metodo Birch, è anche possibile specificare il numero massimo di nodi foglia da creare nell'albero CF.

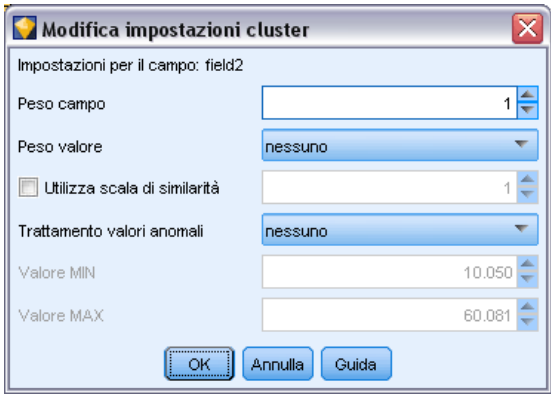
Specifica soglia di similarità. (Solo per il raggruppamento tramite cluster demografici) Il limite inferiore della similarità di due record di dati appartenenti allo stesso cluster. Per esempio, un valore di 0,25 significa che i record con valori simili per il 25% saranno probabilmente assegnati allo stesso cluster. Un valore di 1,0 significa che i record devono essere identici per essere assegnati allo stesso cluster.

Impostazioni campo. Per specificare le opzioni per i singoli campi di input, fare clic sulla riga corrispondente nella colonna Impostazioni della tabella Impostazioni campo e scegliere <Specifica impostazioni>.

Specifica delle impostazioni dei campi per il raggruppamento tramite cluster

Qui è possibile specificare le opzioni per i singoli campi di input.

Figura 5-26
Specifica delle impostazioni cluster per un campo di input



Peso campo. Assegna più o meno peso al campo durante il processo di creazione del modello. Per esempio, se si ritiene che questo campo sia relativamente meno importante per il modello rispetto agli altri campi, diminuirne il peso rispetto agli altri campi.

Peso valore. Assegna più o meno peso a determinati valori del campo. Alcuni valori del campo potrebbero essere più comuni di altri. La coincidenza di valori rari in un campo potrebbe essere più significativa per un cluster della coincidenza di valori frequenti. Per ponderare i valori del campo è possibile scegliere uno dei seguenti metodi (in entrambi i casi, i valori rari hanno un peso notevole, mentre quelli frequenti hanno un peso scarso):

- **Logaritmico.** Assegna un peso a ciascun valore a seconda del logaritmo della sua probabilità nei dati di input.
- **Probabilistico.** Assegna un peso a ciascun valore a seconda della sua probabilità nei dati di input.

Per entrambi i metodi è possibile scegliere anche un'opzione con compensazione per compensare la ponderazione del valore applicata a ogni campo. Se si compensa la ponderazione del valore, l'importanza globale del campo ponderato è uguale a quella di un campo non ponderato, indipendentemente dal numero di valori possibili. La ponderazione compensata influisce solo sull'importanza relativa delle coincidenze nell'insieme dei valori possibili.

Utilizza scala di similarità. Selezionare questa casella per utilizzare la scala di similarità per controllare il calcolo della misurazione di similarità di un campo. La scala di similarità può essere indicata come numero assoluto. La specifica viene considerata solo per i campi numerici attivi. Se non si specifica una scala di similarità, viene utilizzato il valore di default (la metà della deviazione standard). Per ottenere un gran numero di cluster, ridurre la similarità media fra coppie di cluster adottando scale di similarità più piccole per i campi numerici.

Trattamento valori anomali. I valori anomali sono valori al di fuori dell'intervallo dei valori specificati per un campo, definiti da Valore MIN e Valore MAX. È possibile decidere come gestire i valori anomali per questo campo.

- L'impostazione di default, nessuno, significa che per i valori anomali non è previsto alcun trattamento particolare.

- Se si sceglie sostituire con MIN o MAX, i valori di campo inferiori a Valore MIN o superiori a Valore MAX vengono sostituiti rispettivamente con i valori MIN o MAX. In questo caso è possibile impostare i valori di MIN e MAX.
- Se si sceglie trattare come mancante, i valori anomali vengono trattati come valori mancanti e ignorati. In questo caso è possibile impostare i valori di MIN e MAX.

Bayes naive ISW

Bayes naive è un algoritmo molto noto per problemi di classificazione. Il modello viene definito *naïve* perché considera tutte le variabili di previsione proposte come indipendenti l'una dall'altra. Bayes naive è un algoritmo veloce e scalabile in grado di calcolare probabilità condizionali per combinazioni di attributi e per l'attributo obiettivo. Dai dati di addestramento viene stabilita una probabilità indipendente che serve a indicare la verosimiglianza di ciascuna classe obiettivo una volta specificata l'occorrenza di ogni categoria di valore da ogni variabile di input.

L'algoritmo di classificazione Bayes naive ISW è un classificatore probabilistico basato su modelli di probabilità che integrano forti presupposizioni di indipendenza.

Opzioni del modello Bayes naive ISW

Figura 5-27
Scheda Modello del nodo Bayes naive ISW

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 di partizione, questa opzione garantisce che per la creazione del modello verranno utilizzati solo i dati della partizione di addestramento. Per ulteriori informazioni, vedere l'argomento Nodo Partizione in il capitolo 4 in *IBM SPSS Modeler 15 Nodi di input, elaborazione e output*.

Esegui test. Se si seleziona questa opzione, dopo la creazione del modello nella partizione di addestramento verrà eseguito un test IBM InfoSphere Warehouse Data Mining. Ciò comporterà un passaggio sulla partizione di test per stabilire informazioni su qualità del modello, grafici lift e così via.

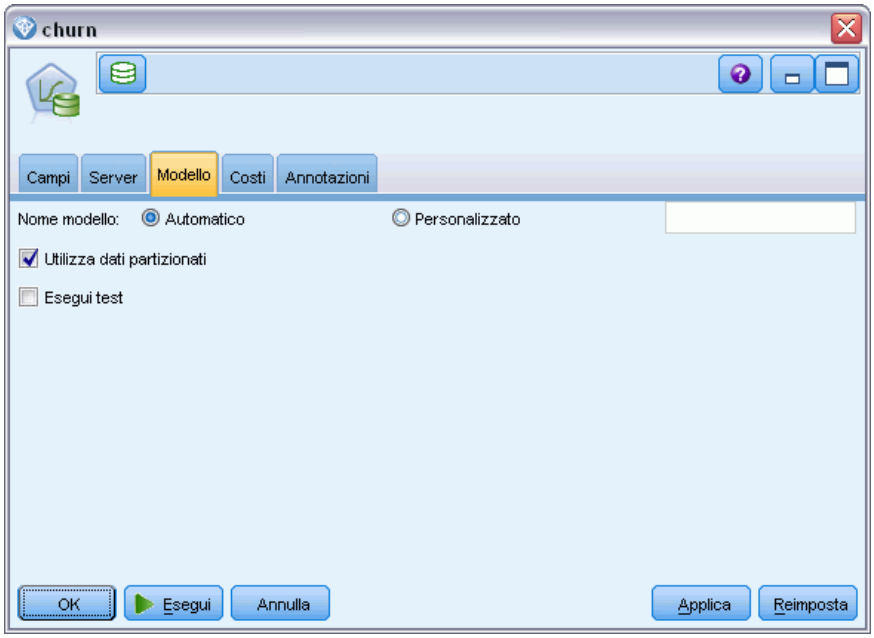
Soglia probabilità. Definisce una probabilità per qualsiasi combinazione di valori predittore e obiettivo non evidenti nei dati di addestramento. Questa probabilità deve essere compresa tra 0 e 1. L'impostazione di default è 0,001.

Regressione logistica ISW

La regressione logistica, nota anche come regressione nominale, è una tecnica statistica per classificare i record in base ai valori dei campi di input. È simile alla regressione lineare, ma l'algoritmo di regressione logistica ISW richiede un campo obiettivo flag (binario) anziché un campo numerico.

Opzioni del modello di Regressione logistica ISW

Figura 5-28
Scheda Modello del nodo Regressione logistica ISW



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 di partizione, questa opzione garantisce che per la creazione del modello verranno utilizzati solo i dati della partizione di addestramento. Per ulteriori informazioni, vedere l'argomento Nodo Partizione in il capitolo 4 in *IBM SPSS Modeler 15 Nodi di input, elaborazione e output*.

Esegui test. Se si seleziona questa opzione, dopo la creazione del modello nella partizione di addestramento verrà eseguito un test IBM InfoSphere Warehouse Data Mining. Ciò comporterà un passaggio sulla partizione di test per stabilire informazioni su qualità del modello, grafici lift e così via.

Serie storica ISW

Gli algoritmi serie storica ISW consentono di prevedere eventi futuri in base a eventi noti verificatisi in passato.

Analogamente ai comuni metodi di regressione, gli algoritmi di serie storica prevedono un valore numerico. A differenza di quanto accade per i comuni metodi di regressione, invece, le previsioni di serie storica si concentrano sui valori futuri di una serie ordinata, definiti in genere previsioni.

Gli algoritmi di serie storica sono univariati, nel senso che la variabile indipendente è una colonna tempo o ordine. Le previsioni si basano su valori passati e non su altre colonne indipendenti.

Gli algoritmi di serie storica sono diversi dai comuni algoritmi di regressione perché non si limitano a prevedere valori futuri ma incorporano nella previsione anche dei cicli stagionali.

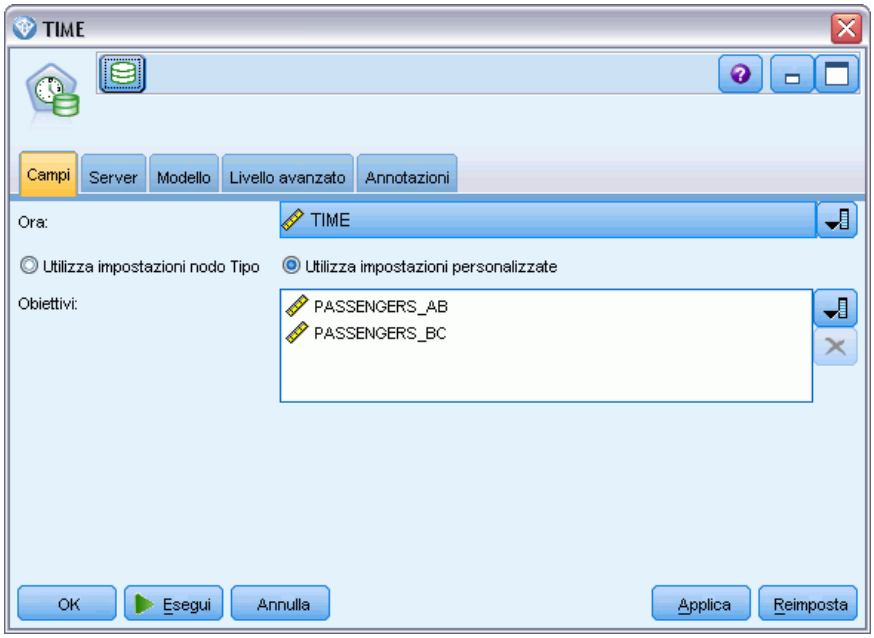
La funzione mining Serie storica dispone dei seguenti algoritmi per la previsione di trend futuri:

- Modello Autoregressivo Integrato a Media Mobile (ARIMA)
- Livellamento esponenziale
- Scomposizione trend stagionale

L'algoritmo che crea la migliore previsione in base ai dati disponibili parte da ipotesi di modello diverse. È possibile calcolare tutte le previsioni contemporaneamente. Gli algoritmi calcolano una previsione dettagliata che comprende il comportamento stagionale della serie storica originale. Se è installato il client IBM InfoSphere Warehouse è possibile utilizzare il Visualizer di serie storiche per valutare e confrontare le curve risultanti.

Opzioni Campi Serie storica ISW

Figura 5-29
Scheda Campi del nodo Serie storica ISW



Ora. Selezionare il campo di input contenente la serie storica. Deve trattarsi di un campo con tipo di archiviazione Data, Ora, Timestamp, Numero reale o Numero intero.

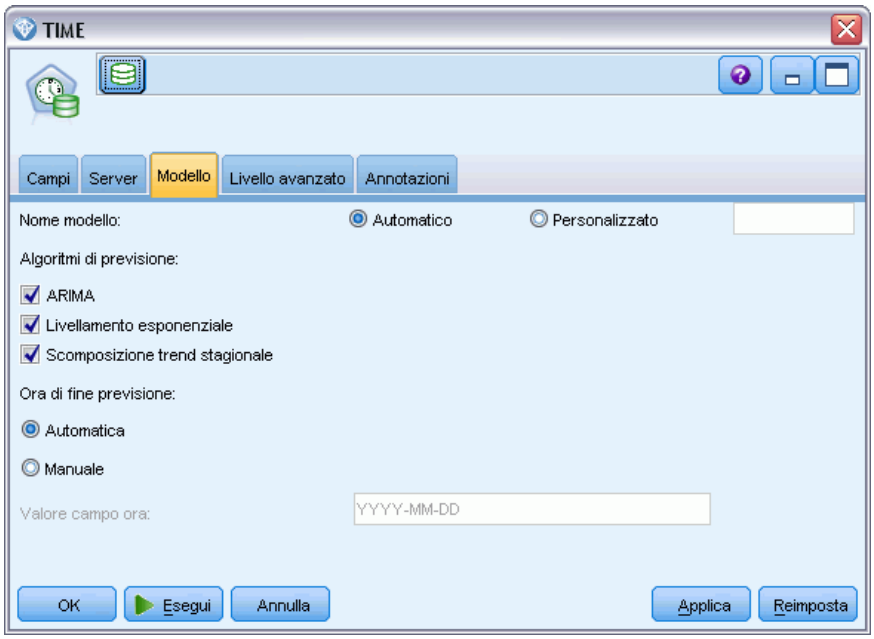
Utilizza impostazioni nodo Tipo. Questa opzione indica al nodo di utilizzare le informazioni sui campi da un nodo Tipo a monte. È l'impostazione di default.

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.

Obiettivi. Selezionare uno o più campi obiettivo. Questa operazione è simile all'impostazione del ruolo di un campo su *Obiettivo* in un nodo Tipo.

Opzioni del modello di serie storica ISW

Figura 5-30
Scheda Modello del nodo Serie storica ISW



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.

Algoritmi di previsione. Selezionare gli algoritmi da utilizzare per la modellazione. È possibile scegliere una o più delle opzioni seguenti:

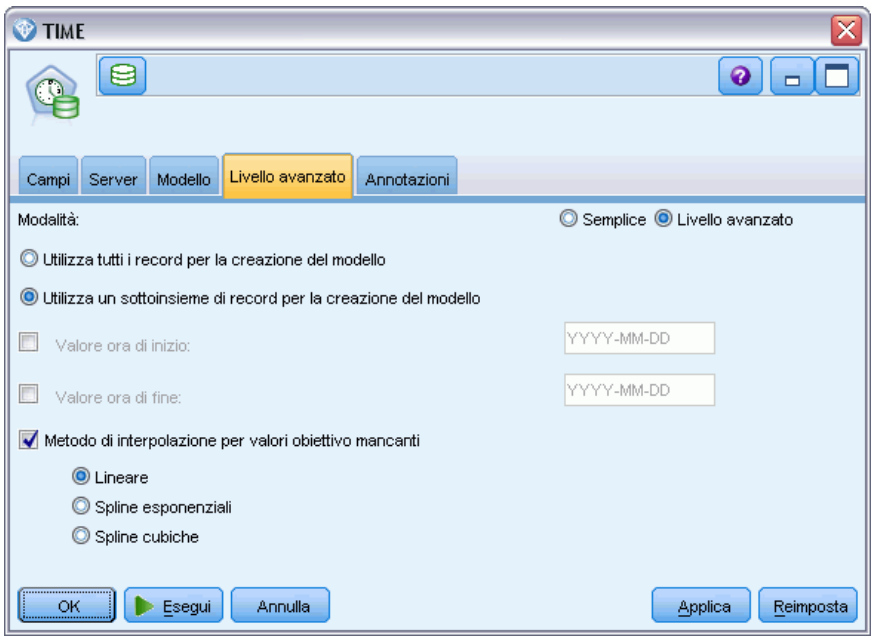
- ARIMA
- Livellamento esponenziale
- Scomposizione trend stagionale.

Ora di fine previsione. Specifica se l'ora di fine della previsione deve essere calcolata automaticamente o indicata manualmente.

Valore campo ora. Quando l'opzione Ora di fine previsione è impostata su manuale, immettere l'ora di fine della previsione. Il valore che è possibile inserire dipende dal tipo del campo Ora; per esempio, se il valore è un integer che rappresenta le ore è possibile inserire 48 per interrompere la previsione dopo l'elaborazione di 48 ore di dati. In alternativa, questo campo può richiedere di immettere una data o un'ora come valore di fine.

Opzioni avanzate per le serie storiche ISW

Figura 5-31
Scheda Livello avanzato del nodo Serie storica ISW



Utilizza tutti i record per la creazione del modello. Questa è l'impostazione di default; quando viene creato il modello, vengono analizzati tutti i record.

Utilizza un sottoinsieme di record per la creazione del modello. Se si desidera creare il modello a partire da solo alcuni dei dati disponibili, selezionare questa opzione. Per esempio, questo potrebbe essere necessario in presenza di un volume eccessivo di dati ripetitivi.

Immettere il Valore ora di inizio e il Valore ora di fine per identificare i dati da utilizzare. Si noti che i valori che è possibile digitare in questi campi dipendono dal tipo del campo Ora, che può essere per esempio un numero di ore o di giorni, oppure una data e un'ora specifica.

Metodo di interpolazione per valori obiettivo mancanti. In caso di elaborazione di dati con uno o più valori mancanti, selezionare il metodo da utilizzare per calcolarli. È possibile scegliere tra le opzioni seguenti:

- Lineare
- Spline esponenziali
- Spline cubiche

Visualizzazione dei modelli di serie storica ISW

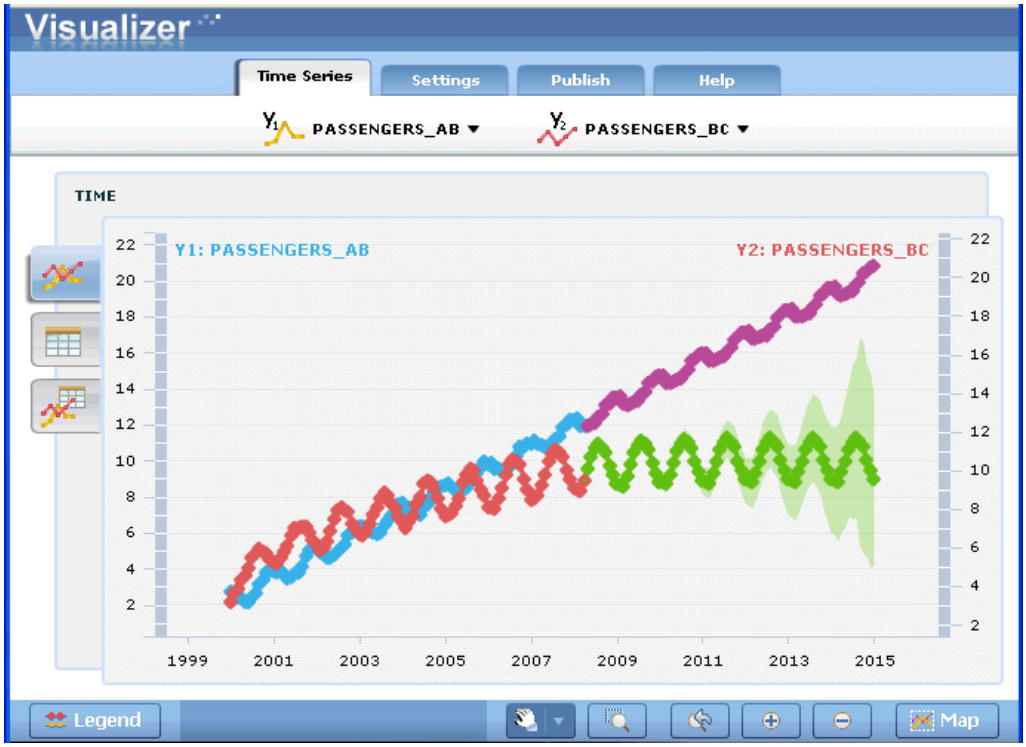
I modelli di serie storica ISW vengono creati sotto forma di modelli grezzi, contenenti informazioni estratte dai dati ma non destinati direttamente alla generazione di previsioni.

Figura 5-32
Icona di modello grezzo

Per ulteriori informazioni, vedere l'argomento Modelli grezzi in il capitolo 3 in *IBM SPSS Modeler 15 Nodi Modelli*.

Se è installato il client IBM InfoSphere Warehouse, è possibile utilizzare il Visualizer di serie storiche per visualizzare una riproduzione grafica dei dati di serie storica.

Figura 5-33
Modello di serie storica ISW visualizzato nel Visualizer



Per utilizzare lo strumento Visualizer di serie storiche:

- Verificare di avere eseguito tutte le operazioni per l'integrazione di IBM® SPSS® Modeler con IBM InfoSphere Warehouse. Per ulteriori informazioni, vedere l'argomento Attivazione dell'integrazione con IBM InfoSphere Warehouse a pag. 107.

- Fare doppio clic sull'icona del modello grezzo nella palette Modelli.

- Nella scheda Server della finestra di dialogo, fare clic sul pulsante Visualizza per richiamare il Visualizer nel browser di default.

Insiemi di modelli di ISW Data Mining

È possibile creare modelli dai nodi Albero decisionale, Associazione, Sequenza, Regressione e Raggruppamento tramite cluster ISW inclusi in IBM® SPSS® Modeler.

Scheda Server dell'insieme di modelli ISW

La scheda Server offre le opzioni che consentono di eseguire controlli di uniformità e avviare lo strumento IBM Visualizer.

Figura 5-34
Scheda Server dell'insieme di modelli ISW

IBM® SPSS® Modeler consente di eseguire un controllo dell'uniformità archiviando una stringa identica con la chiave del modello generato sia nel modello ISW che nel modello di SPSS Modeler. Per eseguire questo tipo di verifica, fare clic sul pulsante Controllo nella scheda Server. Per ulteriori informazioni, vedere l'argomento Gestione dei modelli DB2 a pag. 116.

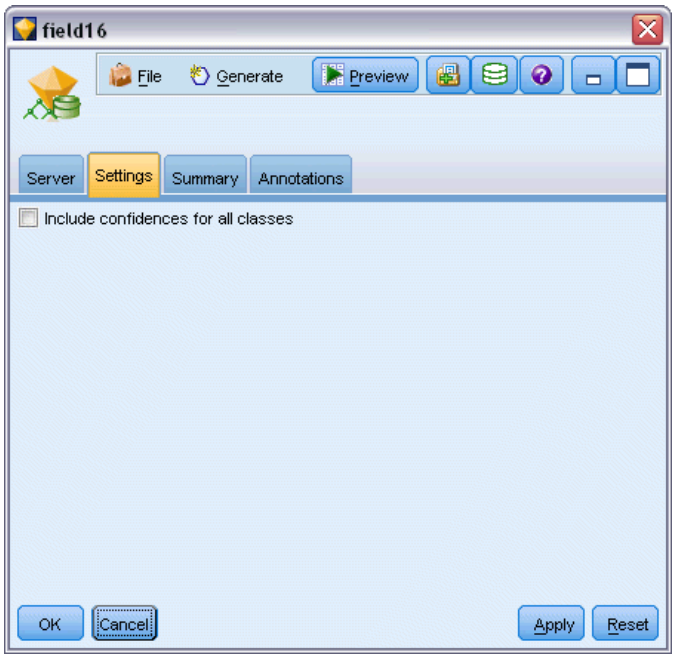
Lo strumento visualizzatore è l'unico modo per sfogliare i modelli di InfoSphere Warehouse Data Mining. Questo strumento può essere installato in via facoltativa con InfoSphere Warehouse Data Mining. Per ulteriori informazioni, vedere l'argomento Attivazione dell'integrazione con IBM InfoSphere Warehouse a pag. 107.

- Fare clic su Visualizza per avviare lo strumento visualizzatore. Ciò che viene visualizzato dallo strumento dipende dal tipo di nodo generato. Per esempio, lo strumento visualizzatore restituirà una visualizzazione Classi previste quando viene avviato da un insieme di modelli Albero decisionale ISW.
- Fare clic su Risultati del test (solo Alberi decisionali e Sequenza) per avviare lo strumento visualizzatore e visualizzare la qualità globale del modello generato.

Scheda Impostazioni dell'insieme di modelli ISW

In IBM® SPSS® Modeler viene generalmente fornita una sola previsione con la probabilità o la confidenza associata. È inoltre disponibile un'opzione utente per la visualizzazione delle probabilità per ogni risultato (simile a quella della regressione logistica) che rappresenta un'opzione tempo punteggio ubicata all'interno della scheda Impostazioni dell'insieme di modelli.

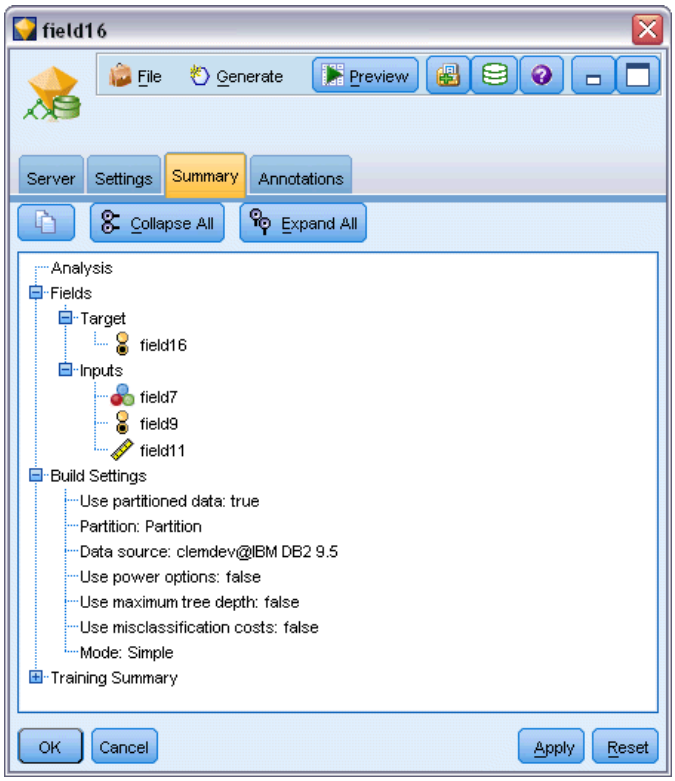
Figura 5-35
Scheda Impostazioni dell'insieme di modelli ISW



Includi confidenze per tutte le classi. Aggiunge una colonna con il livello di confidenza per ciascuno dei possibili risultati del campo obiettivo.

Scheda Riepilogo dell'insieme di modelli ISW

Figura 5-36
Scheda Riepilogo dell'insieme di modelli ISW



La scheda Riepilogo di un insieme di modelli 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 di espansione 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, utilizzare il controllo di espansione di un elemento se si desidera nascondere solo i risultati di tale elemento oppure fare clic sul pulsante Comprimi tutto se si desidera nascondere tutti i risultati.

Analisi. Visualizza informazioni sul modello specifico. Se è stato eseguito un nodo Analisi collegato a questo insieme di modelli, in questa sezione verranno visualizzate anche le informazioni sui risultati di tale analisi. Per ulteriori informazioni, vedere l'argomento nodo Analisi in il capitolo 6 in *IBM SPSS Modeler 15 Nodi di input, elaborazione e output*.

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, lo stream utilizzato per creare tale modello, l'utente che lo ha creato, la data di creazione e il tempo trascorso per la creazione del modello.

Esempi di ISW Data Mining

IBM® SPSS® Modeler per Windows viene fornito con un'ampia gamma di stream di esempio che illustrano il processo di mining in-database. Tali stream si trovano nella cartella di installazione di IBM® SPSS® Modeler in:

\Demos\Database_Modeling\IBM DB2 ISW

Nota: alla cartella Demos è possibile accedere dal gruppo di programmi SPSS Modeler del menu Start di Windows.

Gli stream riportati di seguito possono essere utilizzati insieme, in ordine sequenziale, come esempio del processo di mining in-database:

- *1_upload_data.str*—Utilizzato per la pulitura e il caricamento di dati da un file piatto in DB2.
- *2_explore_data.str*—Utilizzato come esempio di esplorazione dati con SPSS Modeler.
- *3_build_model.str*—Utilizzato per la creazione di un modello Albero decisionale ISW.
- *4_evaluate_model.str*—Utilizzato come esempio di valutazione di modelli con SPSS Modeler.
- *5_deploy_model.str*—Utilizzato per eseguire il deployment del modello ai fini del calcolo del punteggio in-database.

L'insieme di dati impiegato negli stream di esempio concerne applicazioni relative alle carte di credito e presenta un problema di classificazione relativo a un insieme misto di predittori continui e categoriali. Per ulteriori informazioni su questo insieme di dati, fare riferimento al seguente file della cartella di installazione di SPSS Modeler in:

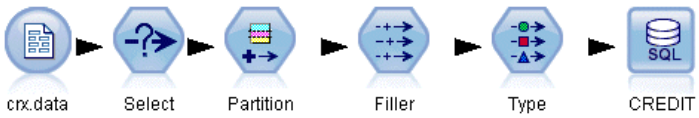
\Demos\Database_Modeling\IBM DB2 ISW\crx.names

L'insieme di dati in questione è disponibile nel repository per l'apprendimento automatico UCI all'indirizzo *http://archive.ics.uci.edu/ml/*.

Stream di esempio: Caricamento dati

Il primo stream di esempio, *1_upload_data.str*, viene utilizzato per pulire e caricare dati da un file piatto in DB2.

Figura 5-37
Stream di esempio usato per il caricamento dei dati

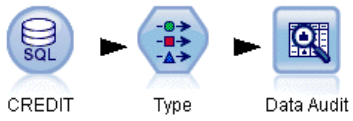


Il nodo Riempimento viene utilizzato per la gestione dei valori mancanti e sostituisce i campi vuoti letti dal file di testo *crx.data* con valori *NULLO*.

Stream di esempio: Explore Data

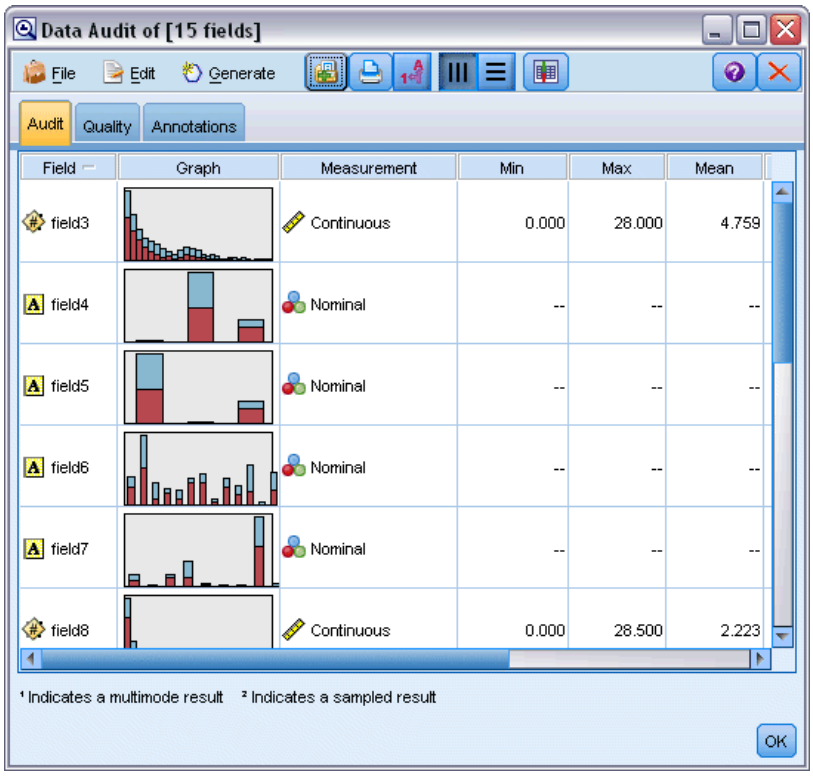
Il secondo stream di esempio, *2_explore_data.str* viene utilizzato per dimostrare l'esplorazione dei dati in IBM® SPSS® Modeler.

Figura 5-38
Stream di esempio usato per esplorare i dati



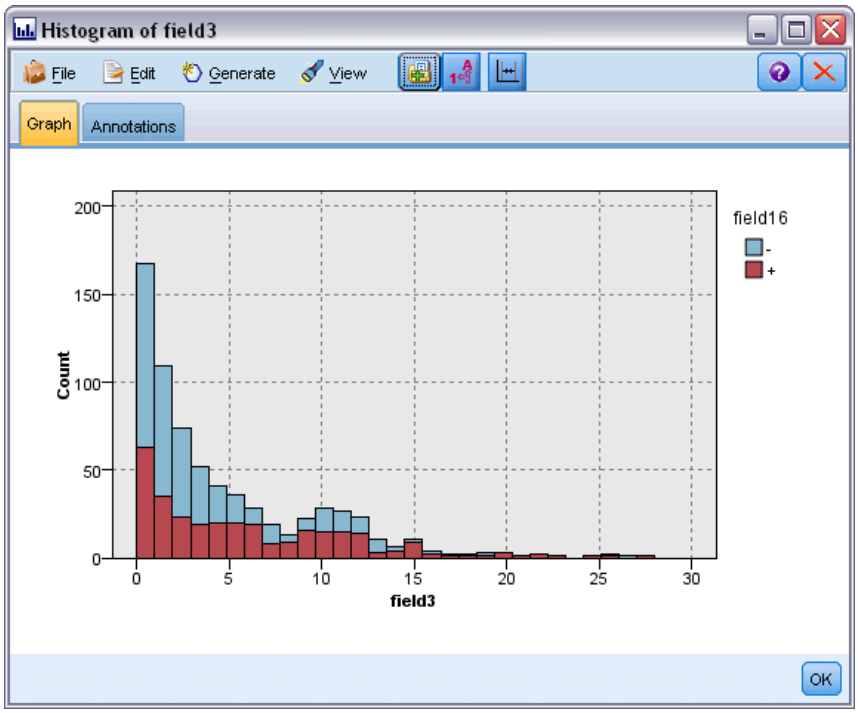
Un passo tipico utilizzato durante l'esplorazione dei dati consiste nell'allegare un nodo Esplora ai dati. Il nodo Esplora è disponibile nella palette di nodi Output.

Figura 5-39
Risultati di Esplora



È possibile utilizzare l'output di un nodo Esplora per acquisire una panoramica generale sui campi e la distribuzione dei dati. Facendo doppio clic su un grafico nella finestra Esplora, si accede a una visualizzazione più dettagliata del grafico che consente un'esplorazione più approfondita di un dato campo.

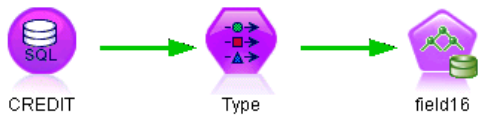
Figura 5-40
Istogramma creato facendo doppio clic su un grafico nella finestra Data Audit



Stream di esempio: Build Model

Il terzo stream di esempio, *3_build_model.str*, illustra la creazione di modelli in IBM® SPSS® Modeler. È possibile collegare il nodo Modelli database allo stream e fare doppio clic per specificare le impostazioni di creazione.

Figura 5-41
Stream di esempio relativo alla modellazione di database, in cui i nodi con ombreggiatura viola indicano l'esecuzione in-database

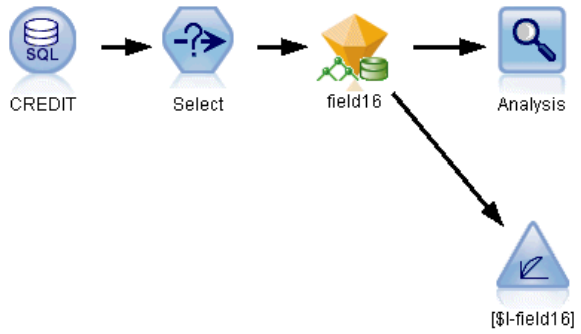


Utilizzando le schede Modello e Livello avanzato del nodo Modelli, è possibile modificare la profondità massima di un albero e bloccare l'ulteriore suddivisione di un nodo dal momento in cui viene creato l'albero decisionale iniziale impostando la massima purezza e il numero minimo di casi per nodo interno. Per ulteriori informazioni, vedere l'argomento Albero decisionale ISW a pag. 122.

Stream di esempio: Valutazione modello

Il quarto stream di esempio, *4_evaluate_model.str*, illustra i vantaggi associati all'utilizzo di IBM® SPSS® Modeler per la modellazione in-database. Una volta eseguito il modello, è possibile aggiungerlo nuovamente allo stream di dati e valutarlo con il supporto di un'ampia gamma di strumenti mirati disponibili in SPSS Modeler.

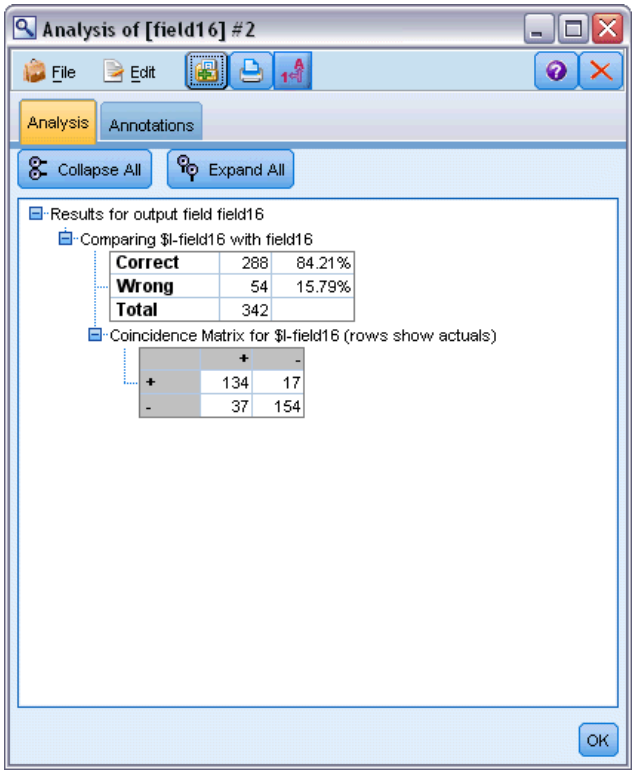
Figura 5-42
Stream di esempio usato per la valutazione del modello



Quando si apre lo stream per la prima volta, l'insieme di modelli (*campo16*) non è compreso nello stream. Aprire il nodo di input CREDIT e verificare di avere specificato una sorgente di dati. Quindi, a condizione di avere eseguito lo stream *3_build_model.str* per creare un insieme di modelli *campo16* nella palette Modelli, è possibile eseguire i nodi disconnessi facendo clic sul pulsante Esegui nella barra degli strumenti (il pulsante contrassegnato da un triangolo verde). In tal modo viene eseguito uno script che copia l'insieme di modelli *campo16* nello stream, quindi lo connette ai nodi esistenti e, infine, esegue i nodi terminali nello stream.

È possibile collegare un nodo Analisi (disponibile nella palette Output) per creare una matrice di coincidenza che mostri lo schema di corrispondenze tra ogni campo generato (previsto) e il relativo campo obiettivo. Eseguire il nodo Analisi per visualizzare i risultati.

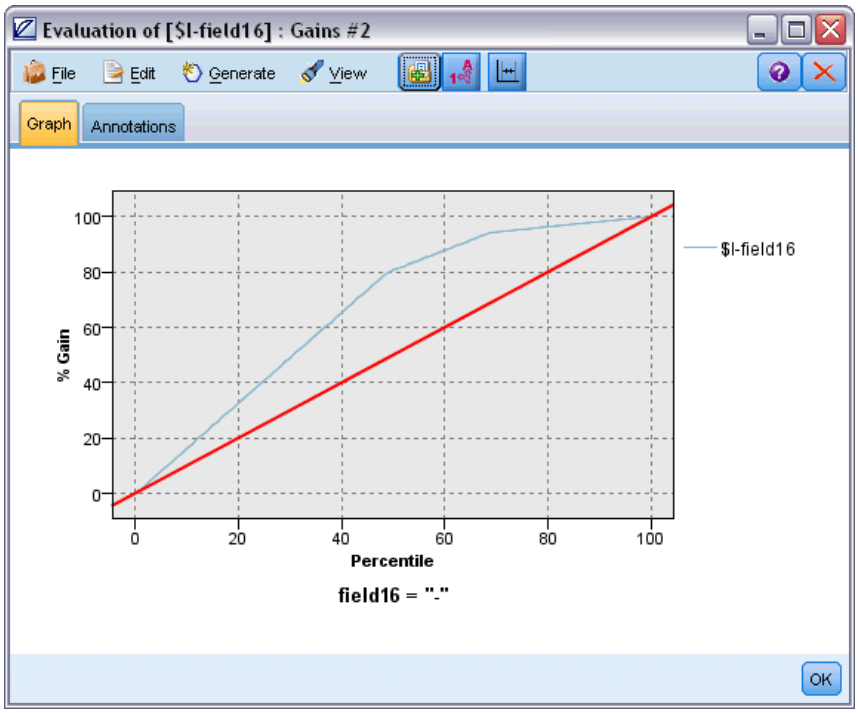
Figura 5-43
Risultati del nodo Analisi



La tabella generata indica che l'84,21% delle previsioni generate dall'algoritmo per alberi decisionali ISW era corretto.

È anche possibile creare un grafico dei guadagni in modo da mostrare i miglioramenti in termini di precisione realizzati dal modello. Collegare un nodo Valutazione al modello generato, quindi eseguire lo stream per visualizzare i risultati.

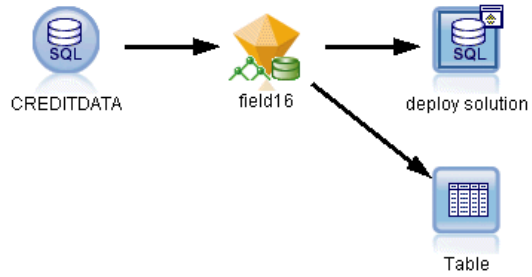
Figura 5-44
Grafico dei guadagni generato mediante il nodo Valutazione



Stream di esempio: Deployment modello

Una volta raggiunto il livello di precisione del modello desiderato, è possibile eseguire il deployment del modello per consentirne l'utilizzo con applicazioni esterne o la riscrittura dei punteggi nel database. Nello stream di esempio *5_deploy_model.str* i dati vengono letti dalla tabella CREDIT. Quando si esegue il nodo di esportazione database *soluzione di deployment*, il punteggio dei dati non viene calcolato. Lo stream crea invece il file di immagine pubblicato *credit_scorer.pim* e il file di parametri pubblicato *credit_scorer.par*.

Figura 5-45
Stream di esempio usato per il deployment del modello



Come nell'esempio precedente, lo stream esegue uno script che copia l'insieme di modelli *campo16* nello stream dalla palette Modelli, lo connette ai nodi esistenti e, infine, esegue i nodi terminali nello stream. In questo caso occorre prima specificare una sorgente di dati sia nei nodi di input Database che nei nodi di esportazione.

Capitolo

6

Modellazione di database con IBM Netezza Analytics

IBM SPSS Modeler e IBM Netezza Analytics

IBM® SPSS® Modeler supporta l'integrazione con IBM® Netezza® Analytics, che consente di eseguire analisi avanzate sui server IBM Netezza. Queste funzionalità sono accessibili tramite l'interfaccia utente grafica e l'ambiente di sviluppo basato sui flussi di lavoro di SPSS Modeler e consentono di eseguire gli algoritmi di data mining direttamente nell'ambiente IBM Netezza.

SPSS Modeler supporta l'integrazione con i seguenti algoritmi di Netezza Analytics.

- Alberi decisionali
- K-Means
- Rete di Bayes
- Bayes naive
- KNN
- Raggruppamento cluster divisivo
- PCA
- Albero di regressione
- Regressione lineare

Per ulteriori informazioni sugli algoritmi vedere la *Netezza Analytics Developer's Guide* e la *Guida di riferimento Netezza Analytics*.

Requisiti per l'integrazione con IBM Netezza Analytics

Di seguito sono elencate le condizioni che costituiscono i prerequisiti indispensabili per l'esecuzione della modellazione in-database con IBM® Netezza® Analytics. Per garantire che queste condizioni vengano soddisfatte, può essere necessario consultare l'amministratore di sistema.

- IBM® SPSS® Modeler eseguito in modalità locale o su un'installazione di IBM® SPSS® Modeler Server su Windows o UNIX (eccetto zLinux per cui non sono disponibili driver ODBC di IBM Netezza).
- IBM Netezza Performance Server 6.0 o versioni successive, con il pacchetto IBM® SPSS® In-Database Analytics.

- Una sorgente dati ODBC per la connessione a un database IBM Netezza. Per ulteriori informazioni, vedere l'argomento Attivazione dell'integrazione con IBM Netezza Analytics a pag. 164.
- Generazione e ottimizzazione SQL abilitate in SPSS Modeler. Per ulteriori informazioni, vedere l'argomento Attivazione dell'integrazione con IBM Netezza Analytics a pag. 164.

Nota: le funzionalità di modellazione in-database e ottimizzazione SQL richiedono che sul computer IBM® SPSS® Modeler sia attivata la connettività SPSS Modeler Server. Con questa impostazione attivata, è possibile accedere agli algoritmi di database, restituire codice SQL direttamente da SPSS Modeler e accedere a SPSS Modeler Server. Per verificare lo stato attuale della licenza, scegliere le seguenti opzioni dal menu SPSS Modeler.

Guida > Informazioni su > Ulteriori dettagli

Se la connettività è abilitata, l'opzioneAbilitazione server viene visualizzata nella scheda Stato della licenza.

Per ulteriori informazioni, vedere l'argomento Connessione a IBM SPSS Modeler Server in il capitolo 3 in *Manuale dell'utente di IBM SPSS Modeler 15*.

Attivazione dell'integrazione con IBM Netezza Analytics

L'attivazione dell'integrazione con IBM® Netezza® Analytics prevede i seguenti passaggi.

- Configurazione di Netezza Analytics
- Creazione di una sorgente ODBC
- Attivazione dell'integrazione in IBM® SPSS® Modeler
- Attivazione della generazione e dell'ottimizzazione SQL in SPSS Modeler

I passaggi sono descritti nelle sezioni che seguono.

Configurazione di IBM Netezza Analytics

Per installare e configurare IBM® Netezza® Analytics vedere la documentazione di Netezza Analytics— in particolare, il documento *Netezza Analytics Installation Guide*—per ulteriori dettagli. La sezione *Setting Database Permissions* di quel manuale contiene informazioni sugli script che è necessario eseguire per consentire agli stream di IBM® SPSS® Modeler di scrivere nel database.

Nota: se si utilizzano nodi che si affidano al calcolo delle matrici (PCA Netezza e Regressione lineare Netezza), è necessario inizializzare il modulo delle matrici di Netezza eseguendo CALL NZM..INITIALIZE(); altrimenti l'esecuzione delle procedure archiviate avrà esito negativo. L'inizializzazione è un passaggio della configurazione da eseguire una sola volta per ogni database.

Creazione di una sorgente ODBC per IBM Netezza Analytics

Per attivare la connessione tra il database IBM Netezza e IBM® SPSS® Modeler è necessario creare un nome di sorgente dati (DSN, Data Source Name) ODBC.

Per creare un DSN, è necessario avere una conoscenza di base delle sorgenti dati e dei driver ODBC e disporre del supporto database in SPSS Modeler. Per ulteriori informazioni, vedere l'argomento Accesso ai dati in il capitolo 2 in *IBM SPSS Modeler Server 15 Guida della performance e amministrazione*.

Se l'applicazione è in esecuzione in modalità distribuita su IBM® SPSS® Modeler Server, creare il DSN sul computer server. Se invece è attiva la modalità locale (client), creare il DSN sul computer client.

Client Windows

- Dal CD del *client Netezza* eseguire il file *nzodbcsetup.exe* per avviare il programma di installazione. Attenersi alle istruzioni visualizzate per installare il driver. Per le istruzioni complete, vedere *IBM Netezza ODBC, JDBC, and OLE DB Installation and Configuration Guide*.
- Creare il DSN.

  Nota: la sequenza dei menu dipende dalla versione di Windows in uso.

  - **Windows XP.** Dal menu Start, scegliere Pannello di controllo. Fare doppio clic su Strumenti di amministrazione, quindi ancora doppio clic su Origini dati (ODBC).
  - **Windows Vista.** Dal menu Start, scegliere Pannello di controllo, quindi Strumenti di amministrazione. Fare doppio clic su Strumenti di amministrazione, selezionare Origini dati (ODBC) quindi Apri.
  - **In Windows 7.** Dal menu Start, scegliere Pannello di controllo, quindi Sistema e sicurezza e Strumenti di amministrazione. Selezionare Origini dati (ODBC) e fare clic su Apri.
- Fare clic sulla scheda DSN di sistema, quindi fare clic su Aggiungi.
- Selezionare NetezzaSQL dall'elenco e fare clic su Fine.
- Nella scheda delle opzioni DSN della schermata IBM Netezza ODBC Driver Setup, digitare un nome di sorgente dati, il nome host o indirizzo IP del server IBM Netezza, il numero di porta per la connessione, il database dell'istanza Netezza in uso e il nome utente e la password utilizzati per la connessione al database. Fare clic sul pulsante Guida per visualizzare una spiegazione dei campi.
- Fare clic sul pulsante Verifica connessione e assicurarsi di poter eseguire la connessione al database.
- Stabilita correttamente la connessione, fare clic su OK più volte per uscire dalla schermata Amministratore origine dati ODBC.

Server Windows

La procedura per Windows Server è uguale alla procedura per il client in Windows XP.

Server UNIX o Linux

La procedura che segue è valida per i server UNIX o Linux (tranne zLinux, per il quale non sono disponibili driver ODBC di IBM Netezza).

- Dal CD del *client Netezza* copiare il file *<platform>cli.package.tar.gz* pertinente in una posizione temporanea sul server.

- Estrarre i contenuti dell'archivio utilizzando i comandi gunzip e untar.
- Aggiungere le autorizzazioni per l'esecuzione allo script *unpack* estratto.
- Eseguire lo script, rispondendo ai prompt visualizzati.
- Modificare il file *modelersrv.sh* in modo che includa le righe riportate di seguito.

```
. /usr/IBM/SPSS/SDAP61_notfinal/odbc.sh
LD_LIBRARY_PATH_64=$LD_LIBRARY_PATH:/usr/local/nz/lib64; export LD_LIBRARY_PATH_64
NZ_ODBC_INI_PATH=/usr/IBM/SPSS/SDAP61_notfinal; export NZ_ODBC_INI_PATH
```

- Individuare il file */usr/local/nz/lib64/odbc.ini* e copiarne il contenuto nel file *odbc.ini* installato con SDAP 6.1 (quello definito dalla variabile di ambiente $ODBCINI).

  Nota: per i sistemi Linux a 64 bit, il parametro *Driver* fa riferimento per errore al driver a 32 bit. Quando si copia il contenuto di *odbc.ini* nel passaggio precedente, modificare il percorso in questo parametro come nell'esempio che segue:

  ```
  /usr/local/nz/lib64/libnzodbc.so
  ```

- Modificare i parametri nella definizione DSN Netezza in modo che riflettano il database da utilizzare.
- Riavviare SPSS Modeler Server e provare a utilizzare i nodi di mining in-database di Netezza sul client.

Attivazione dell'integrazione IBM Netezza Analytics in IBM SPSS Modeler

- Dal menu principale IBM® SPSS® Modeler, scegliere
  Strumenti > Opzioni > Applicazioni di supporto.
- Fare clic sulla scheda IBM Netezza.

  Attiva integrazione di Netezza Data Mining. Attiva la palette Modelli in-database (se non è già visualizzata) nella parte inferiore della finestra SPSS Modeler e aggiunge i nodi degli algoritmi di Netezza Data Mining.

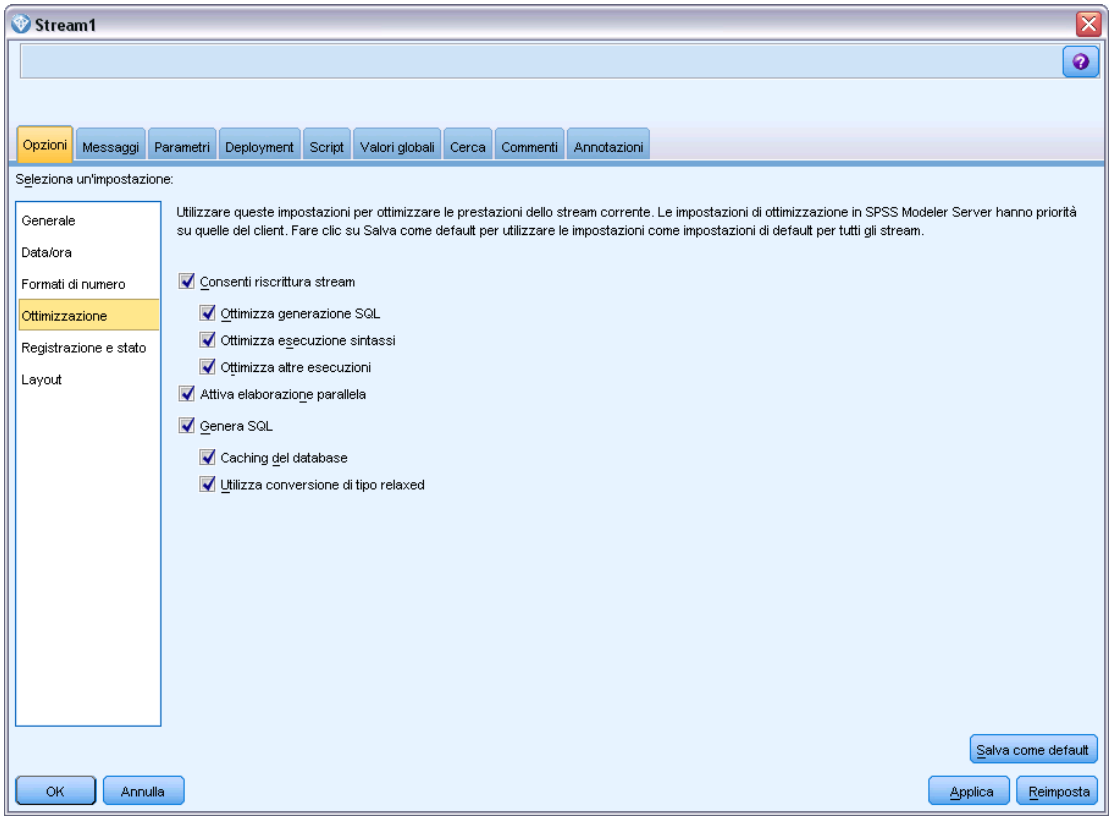
  Connessione Netezza. Fare clic sul pulsante Modifica e scegliere la stringa di connessione Netezza specificata al momento della creazione della sorgente ODBC. Per ulteriori informazioni, vedere l'argomento Creazione di una sorgente ODBC per IBM Netezza Analytics a pag. 164.

Attivazione di generazione e ottimizzazione SQL

Poiché è probabile che ci si trovi a lavorare con insiemi di dati di dimensioni molto grandi, per motivi di prestazioni è bene attivare le opzioni di generazione e ottimizzazione SQL in IBM® SPSS® Modeler.

- Dai menu di SPSS Modeler scegliere:
  Strumenti > Proprietà stream > Opzioni

Figura 6-1
Impostazioni di ottimizzazione



- Fare clic sull'opzione Ottimizzazione nel riquadro di spostamento.
- Confermare che l'opzione Genera SQL è attivata. Questa impostazione è necessaria per il corretto funzionamento della modellazione di database.
- Selezionare Ottimizza generazione SQL e Ottimizza altre esecuzioni (queste due opzioni non sono strettamente necessarie, tuttavia se ne consiglia la selezione per ottenere performance ottimizzate).

  Per ulteriori informazioni, vedere l'argomento Impostazione delle opzioni di ottimizzazione per gli stream in il capitolo 5 in *Manuale dell'utente di IBM SPSS Modeler 15*.

Creazione di modelli con IBM Netezza Analytics

Per ognuno degli algoritmi supportati esiste un nodo Modelli corrispondente. Ai nodi Modelli di IBM Netezza è possibile accedere dalla scheda Modelli database nella palette dei nodi. Per ulteriori informazioni, vedere l'argomento Palette nodi in il capitolo 3 in *Manuale dell'utente di IBM SPSS Modeler 15*.

Considerazioni sui dati

I campi della sorgente dati possono contenere variabili di diversi tipi di dati, in base al nodo Modelli. In IBM® SPSS® Modeler, i tipi di dati sono noti come **livelli di misurazione**. La scheda Campi del nodo Modelli utilizza delle icone per indicare i tipi di livello di misurazione consentiti per i campi di input e obiettivo. Per ulteriori informazioni, vedere l'argomento Livelli di misurazione in il capitolo 4 in *IBM SPSS Modeler 15 Nodi di input, elaborazione e output*.

Campo obiettivo. Il campo obiettivo è il campo il cui valore si tenta di prevedere. Dove può essere specificato un obiettivo, è possibile selezionare come campo obiettivo un solo campo dati di origine.

Campo ID record. Specifica il campo utilizzato per identificare ciascun caso in modo univoco. Tale campo può coincidere, per esempio, con un campo ID quale *IDCliente*. Se i dati di origine non includono un campo ID, è possibile crearlo mediante un nodo Nuovo campo, come indica la procedura che segue.

- Selezionare il nodo di origine.
- Nella scheda Oper su campi della palette dei nodi, fare doppio clic sul nodo Nuovo campo.
- Aprire il nodo Nuovo campo facendo doppio clic sulla relativa icona nell'area di disegno.
- Nel campo Nuovo campo digitare, per esempio, ID.
- Nel campo Formula digitare @INDEX e fare clic su OK.
- Collegare il nodo Nuovo campo al resto dello stream.

Gestione dei valori nulli

Se i dati di input contengono valori nulli, l'utilizzo di alcuni nodi Netezza potrebbe causare messaggi di errore o stream lunghi da eseguire, quindi è consigliabile rimuovere i record che contengono valori nulli. Utilizzare il metodo seguente.

- Collegare un nodo Seleziona al nodo di input.
- Impostare l'opzione Modalità del nodo Seleziona su Scarta.
- Immettere quanto segue nel campo Condizione:

  @NULL(*field1*) [or @NULL(*field2*)[... or @NULL(*fieldN*]])

  Assicurarsi di includere tutti i campi di input.
- Collegare il nodo Seleziona al resto dello stream.

Output modello

È possibile che uno stream contenente un nodo di modellazione Netezza produca risultati leggermente diversi a ogni esecuzione. Questo si verifica perché l'ordine con il quale il nodo legge i dati di input non è sempre lo stesso poiché i dati vengono letti in tabelle temporanee prima della creazione dei modelli. Le differenze prodotte da questo effetto sono tuttavia trascurabili.

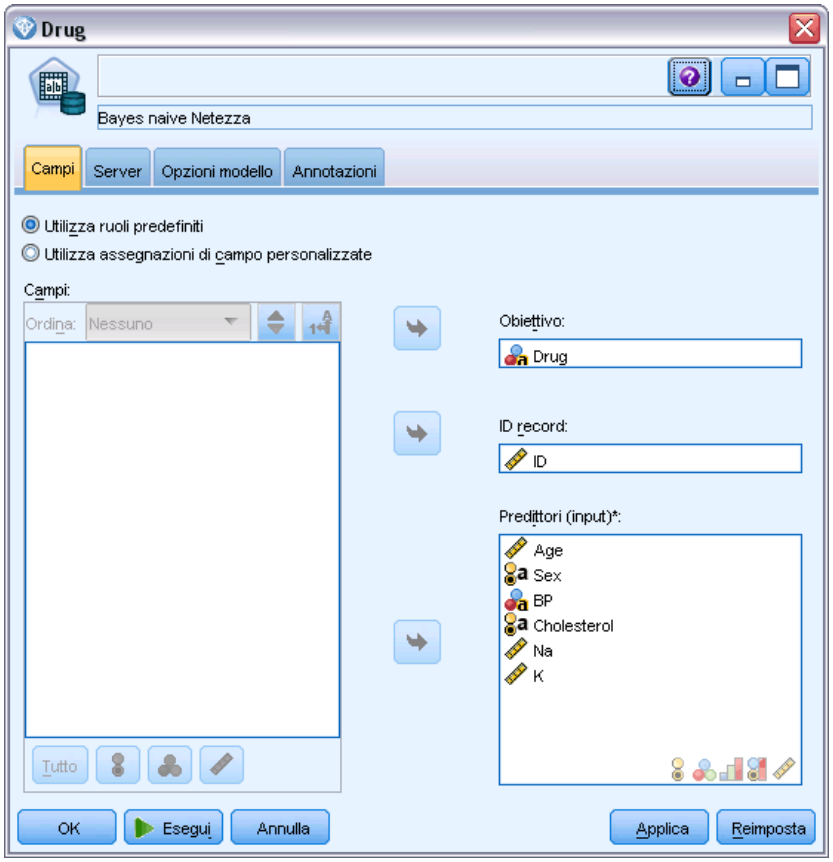
Commenti generali

- In IBM® SPSS® Collaboration and Deployment Services non è possibile creare configurazioni per il calcolo del punteggio utilizzando stream che contengono nodi Modelli database IBM Netezza.
- Per i modelli creati dai nodi Netezza non è possibile eseguire l'esportazione o l'importazione PMML.

Opzioni della scheda Campi dei modelli Netezza

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.

Figura 6-2
Esempio di opzioni dei campi di Netezza



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 di origine a monte. Per ulteriori informazioni, vedere l'argomento Impostazione del ruolo del campo in il capitolo 4 in *IBM SPSS Modeler 15 Nodi di input, elaborazione e output*.

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. Per i modelli lineari generalizzati, vedere anche il campo Prove di questo schermo.

ID record. Campo da utilizzare come identificatore univoco del record.

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

Opzioni della scheda Server dei modelli Netezza

In questa scheda è possibile specificare il database IBM Netezza in cui deve essere archiviato il modello.

Figura 6-3
Esempio di opzioni del server di Netezza

Dettagli Server DB Netezza. Qui si specificano i dettagli della connessione per il database da utilizzare per il modello.

- **Utilizza connessione a monte**. (default) Utilizza i dettagli di connessione specificati in un nodo a monte, per esempio il nodo di input Database. *Nota*: questa opzione funziona solo se tutti i nodi a monte sono in grado di utilizzare il push back SQL. In questo caso non è necessario spostare i dati fuori dal database perché SQL supporta pienamente tutti i nodi a monte.
- **Sposta dati alla connessione**. Sposta i dati nel database indicato qui. In questo modo si consente al modello di lavorare se i dati si trovano su un altro database IBM Netezza, o su un database di un altro fornitore, o anche se i dati si trovano in un file piatto. Inoltre i dati vengono riportati in questo database se sono stati in precedenza estratti perché un nodo non ha effettuato il push back SQL. Fare clic sul pulsante Modifica per reperire e selezionare una connessione. *Attenzione*: IBM® Netezza® Analytics è utilizzato generalmente con insiemi di dati molto grandi. Il trasferimento di grandi quantità di dati tra database, o anche dentro e fuori lo stesso database, può richiedere molto tempo ed è quindi da evitare se possibile.

Nome tabella. Nome della tabella di database in cui verrà archiviato il modello. *Nota*: deve essere una tabella nuova, non è possibile utilizzare una tabella già esistente per questa operazione.

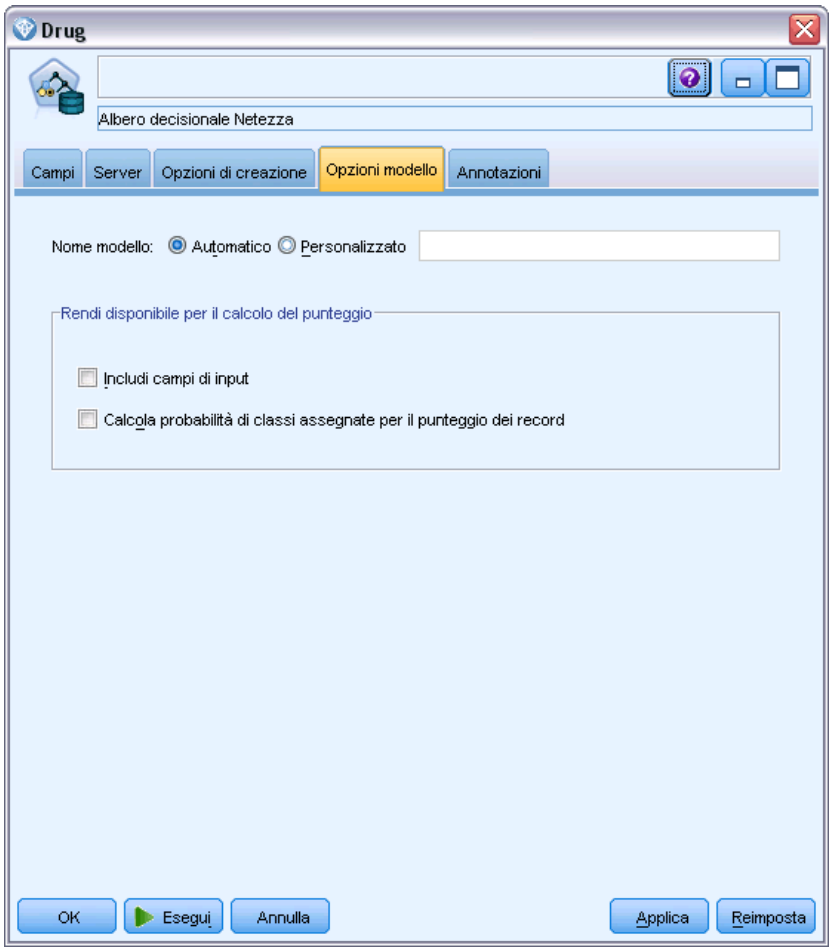
Commenti

- La connessione utilizzata per la modellazione non deve necessariamente corrispondere a quella impiegata nel nodo di input di uno stream. Per esempio, è possibile utilizzare uno stream che accede ai dati di un database IBM Netezza, li scarica in IBM® SPSS® Modeler per la pulitura o altre operazioni di modifica e, infine, li carica in un database IBM Netezza differente per la modellazione. Si noti tuttavia che questa configurazione può avere effetti negativi sulle prestazioni.
- Il nome della sorgente dati ODBC è efficacemente incorporato in ogni stream di SPSS Modeler. Se uno stream creato su un determinato host viene eseguito su un host differente, il nome della sorgente dati deve essere lo stesso su entrambi gli host. In alternativa, è possibile selezionare una sorgente dati differente nella scheda Server all'interno di ogni nodo Modelli o di input.

Modelli Netezza - Opzioni Modello

Nella scheda Opzioni modello è possibile scegliere se specificare un nome per il modello o generare un nome automaticamente. È inoltre possibile impostare valori di default per le opzioni di calcolo del punteggio.

Figura 6-4
Esempio di opzioni dei modelli di Netezza



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.

Rendi disponibile per il calcolo del punteggio. È possibile impostare qui i valori di default per le opzioni di calcolo del punteggio che appaiono nella finestra di dialogo dell'insieme di modelli. Per maggiori dettagli sulle opzioni, vedere l'argomento della Guida relativo alla scheda Impostazioni dell'insieme di modelli specifico.

Alberi decisionali di Netezza

L'albero decisionale è una struttura gerarchica che rappresenta un modello di classificazione. Con un modello di albero decisionale, è possibile sviluppare un sistema di classificazione per prevedere o classificare future osservazioni provenienti da un insieme di dati di addestramento. La classificazione assume l'aspetto di una struttura ad albero in cui i rami rappresentano i punti di suddivisione nella classificazione. In tali punti i dati vengono suddivisi in sottogruppi in modo

ricorsivo finché non viene raggiunto un punto di arresto. I nodi dell'albero sono punti di arresto noti come **foglie**. Ogni foglia assegna un'etichetta, detta **etichetta di classe**, ai membri del relativo sottogruppo (classe).

L'output dei modelli assume la forma di una rappresentazione di testo dell'albero. Ogni riga di testo corrisponde a un nodo o una foglia e il rientro riflette il livello dell'albero. Per un nodo, viene visualizzata la condizione di suddivisione, per una foglia appare l'etichetta di classe assegnata.

Pesi delle istanze e delle classi

Per default, si presume che tutti i record di input e tutte le classi abbiano uguale importanza relativa. Questa impostazione si può modificare assegnando pesi individuali ai membri di uno di questi elementi o di entrambi. Questo può essere utile, per esempio, se i punti dei dati di addestramento non sono distribuiti in modo realistico tra le categorie. I pesi consentono di applicare una distorsione al modello in modo da compensare le categorie meno rappresentate nei dati. L'incremento del peso di un valore di destinazione dovrebbe aumentare la percentuale di previsioni corrette per quella categoria.

Nel nodo Modelli Albero decisionale è possibile specificare due tipi di pesi. I **pesi delle istanze** assegnano un peso a ogni riga dei dati di input. Per la maggior parte dei casi, il peso indicato è in genere 1,0, con valori più alti o più bassi assegnati solo ai casi più o meno importanti rispetto alla maggioranza dei casi, per esempio:

ID record	Obiettivo	Peso istanza.
1	drugA	1.1
2	drugB	1.0
3	drugA	1.0
4	drugB	0.3

I **pesi delle classi** assegnano un peso a ogni categoria del campo obiettivo, per esempio:

Class	Peso della classe
drugA	1.0
drugB	1.5

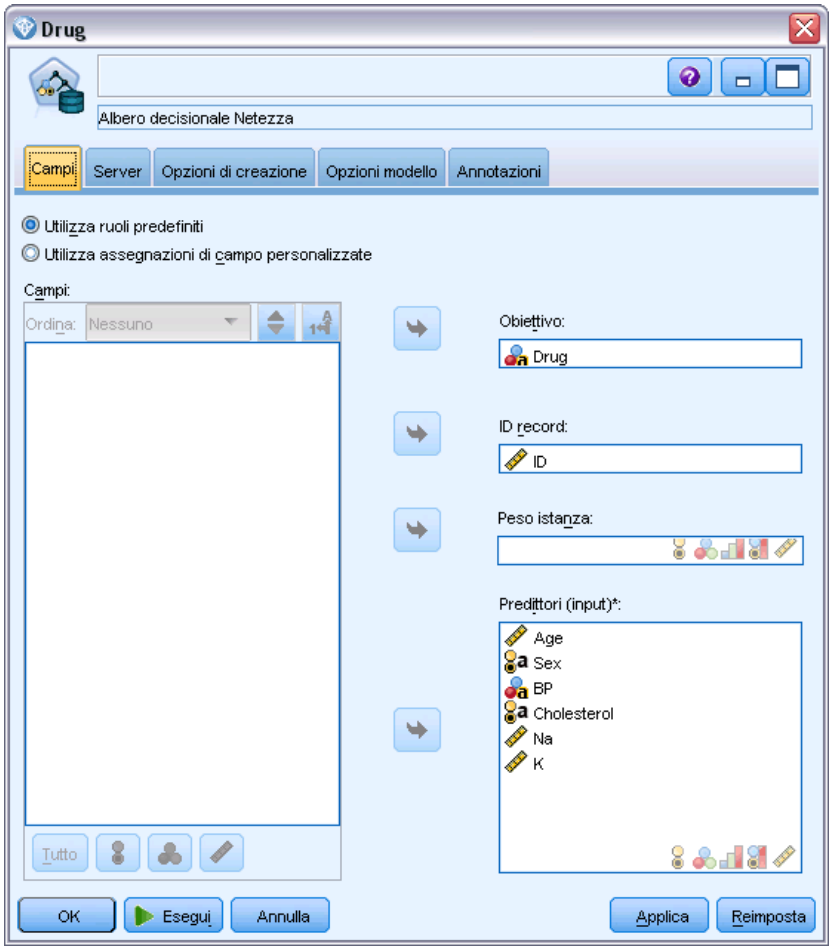
È possibile utilizzare contemporaneamente entrambi i tipi di pesi, nel qual caso essi vengono moltiplicati insieme e utilizzati come peso delle istanze. Quindi, se si utilizzassero insieme i due esempi precedenti, l'algoritmo utilizzerebbe i pesi delle istanze seguenti.

ID record	Calcolo	Peso istanza.
1	1.1*1.0	1.1
2	1.0*1.5	1.5
3	1.0*1.0	1.0
4	0.3*1.5	0.45

Opzioni dei campi dell'albero decisionale di Netezza

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.

Figura 6-5
Opzioni della scheda Campi dell'albero decisionale



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 di origine a monte. Per ulteriori informazioni, vedere l'argomento Impostazione del ruolo del campo in il capitolo 4 in *IBM SPSS Modeler 15 Nodi di input, elaborazione e output*.

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.

ID record. Campo da utilizzare come identificatore univoco del record. I valori di questo campo devono essere univoci per ogni record (per esempio, i numeri di ID dei clienti).

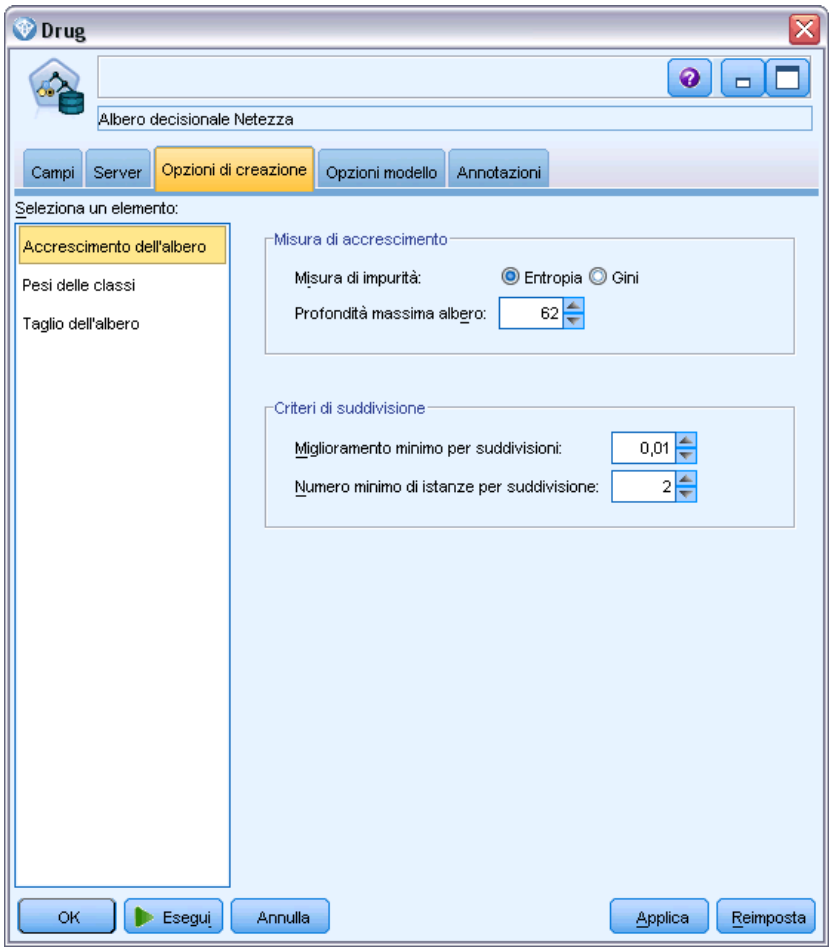
Peso istanza. Se si specifica un campo, è possibile utilizzare i pesi delle istanze (un peso per ogni riga di dati di input) in aggiunta ai pesi delle classi di default (un peso per ogni categoria per il campo obiettivo) o al posto degli stessi. Il campo da specificare qui deve contenere un peso numerico per ogni riga dei dati di input. Per ulteriori informazioni, vedere l'argomento Pesi delle istanze e delle classi a pag. 173.

Predittori (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.

Opzioni di creazione dell'albero decisionale di Netezza

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.

Figura 6-6
Opzioni di creazione dell'albero decisionale per l'accrescimento dell'albero



È possibile impostare le opzioni di creazione per:

- Accrescimento dell'albero
- Pesi per le etichette di classe
- Taglio dell'albero

Le opzioni per l'accrescimento dell'albero sono descritte in questa sezione.

Misura di accrescimento. Queste opzioni controllano il modo in cui viene misurato l'accrescimento dell'albero. Se non si desidera utilizzare i valori di default, fare clic su Personalizza e apportare le modifiche.

- **Misura di impurità.** La misurazione dell'impurità utilizzata per valutare il punto migliore in cui suddividere l'albero. **Impurità** si riferisce al grado di presenza, nei sottogruppi definiti dall'albero, di un ampio intervallo di valori di campi di output all'interno di ciascun gruppo.

Le misure supportate sono **Entropia** (default) e **Gini**, due tipi di misurazione tra i più diffusi, basati sulle probabilità di appartenenza alle categorie per il ramo.

- **Profondità massima albero.** Numero massimo di foglie fino al quale l'albero può crescere sotto il nodo principale, ovvero il numero di volte che il campione può essere suddiviso in modo ricorsivo. Il valore di default è 62, che è la massima profondità possibile dell'albero ai fini della modellazione. Si noti tuttavia che il visualizzatore nell'insieme di modelli è in grado di visualizzare al massimo 10 foglie.

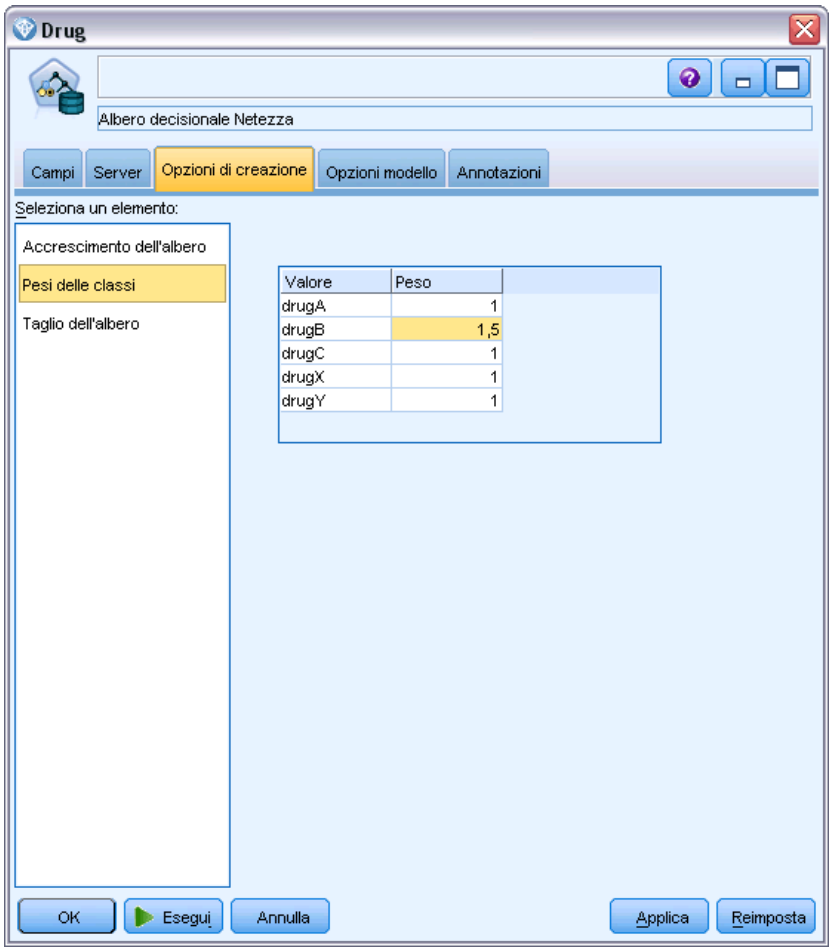
Criteri di suddivisione. Queste opzioni controllano il momento in cui interrompere la suddivisione dell'albero. Se non si desidera utilizzare i valori di default, fare clic su Personalizza e apportare le modifiche.

- **Miglioramento minimo per suddivisioni**. La quantità minima in base alla quale l'impurità deve essere ridotta prima di creare una nuova suddivisione nell'albero. La creazione dell'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 di un valore inferiore a quello specificato dal criterio di suddivisione, la suddivisione non verrà eseguita.
- **Numero minimo di istanze per suddivisione**. Numero minimo di record che possono essere suddivisi. Quando la quantità di record ancora da suddividere è inferiore a questo numero, non verranno eseguite altre suddivisioni. È possibile utilizzare questo campo per impedire che vengano creati sottogruppi molto piccoli nell'albero.

Nodo dell'albero decisionale di Netezza - Pesi delle classi

È possibile assegnare pesi alle singole classi. L'impostazione di default è assegnare il valore 1 a tutte le classi in modo che abbiano lo stesso peso. Specificando valori numerici diversi per i pesi delle singole etichette di classe, si indica all'algoritmo di pesare gli insiemi di addestramento delle singole classi.

Figura 6-7
Opzioni di peso delle classi dell'albero decisionale



Per modificare un peso, farvi doppio clic sopra nella colonna Peso e apportare le modifiche desiderate.

Valore. L'insieme delle etichette di classe ricavato dai valori possibili del campo obiettivo.

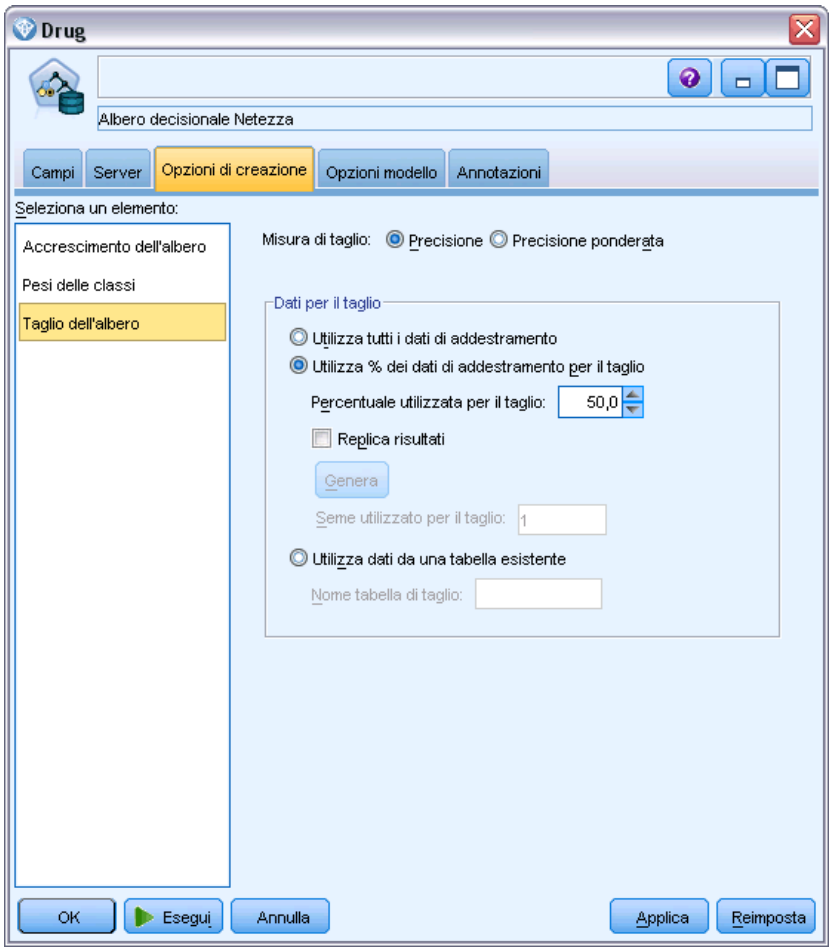
Peso. Il peso da assegnare a una determinata classe. L'assegnazione di un peso superiore a una classe rende il modello più sensibile a quella classe rispetto alle altre.

I pesi delle classi si possono utilizzare insieme ai pesi delle istanze. Per ulteriori informazioni, vedere l'argomento Pesi delle istanze e delle classi a pag. 173.

Nodo dell'albero decisionale di Netezza - Taglio dell'albero

Le opzioni di taglio consentono di specificare i criteri con cui l'albero decisionale viene tagliato. Lo scopo del taglio è ridurre il rischio di sovradattamento rimuovendo i sottogruppi cresciuti troppo che non migliorano la precisione attesa nei nuovi dati.

Figura 6-8
Opzioni di taglio dell'albero decisionale



Misura di taglio. La misura di default del taglio, Precisione, garantisce che la precisione stimata del modello rimanga entro limiti accettabili dopo la rimozione di una foglia dall'albero. Utilizzare invece Precisione ponderata, se si desidera tenere in considerazione i pesi delle classi quando si applica il taglio.

Dati per il taglio. È possibile utilizzare alcuni o tutti i dati di addestramento per stimare la precisione attesa dei nuovi dati. In alternativa, è possibile utilizzare un insieme di dati di taglio separato estratti da una tabella specifica.

- **Utilizza tutti i dati di addestramento**. Questa opzione (default) utilizza tutti i dati di addestramento per stimare la precisione del modello.
- **Utilizza % dei dati di addestramento per il taglio**. Utilizzare questa opzione per dividere i dati in due insiemi, uno per l'addestramento e uno per il taglio, usando la percentuale qui specificata per i dati del taglio.

Selezionare Replica risultati se si desidera specificare un seme aleatorio per assicurarsi che i dati vengano partizionati nello stesso modo ogni volta che si esegue lo stream. È possibile specificare un valore intero nel campo Seme utilizzato per il taglio oppure fare clic su Genera per creare un intero pseudocasuale.

- **Utilizza dati da una tabella esistente**. Specificare il nome della tabella di un insieme di dati di taglio separato per la stima della precisione del modello. Questa operazione è più affidabile rispetto all'utilizzo dei dati di addestramento. Tuttavia, questa opzione può causare la rimozione di un grande sottoinsieme di dati dall'insieme di addestramento, riducendo la qualità dell'albero decisionale.

K-Means Netezza

Il nodo K-Means implementa l'algoritmo *k*-means, che fornisce un metodo di analisi dei cluster. Questo nodo si può utilizzare per raggruppare un insieme di dati in gruppi distinti.

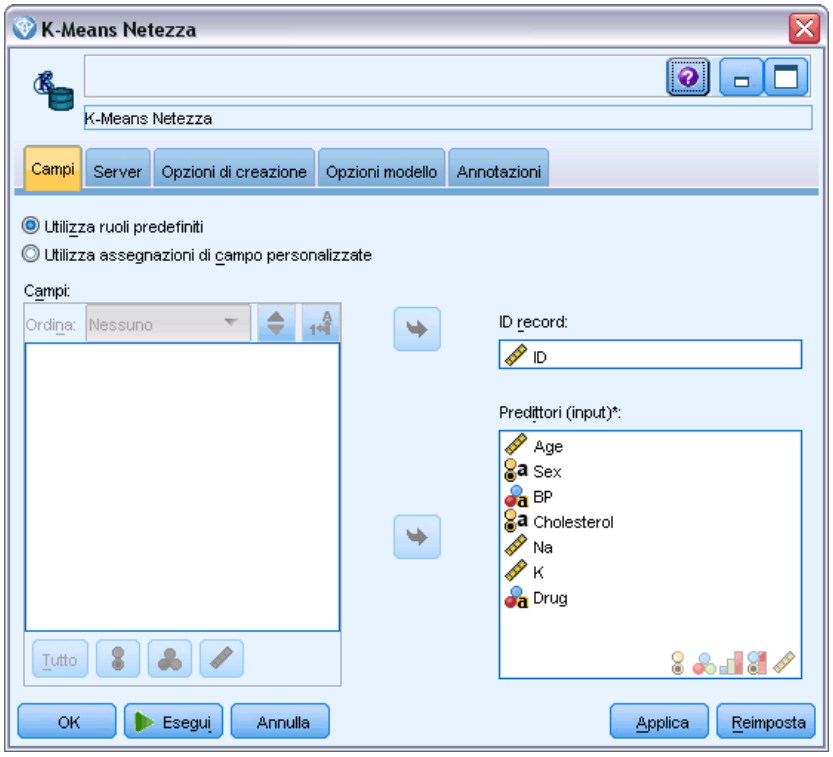
Si tratta di un algoritmo di cluster basato sulla distanza che utilizza una metrica di distanza (funzione) per misurare la similarità fra i punti dei dati. I punti dei dati vengono assegnati al cluster più vicino in base alla metrica di distanza utilizzata.

L'algoritmo funziona eseguendo una serie di iterazioni del medesimo processo di base in cui ogni istanza di addestramento viene assegnata al cluster più vicino (rispetto alla funzione di distanza specificata, applicata all'istanza e al centro del cluster). Tutti i centri dei cluster vengono in seguito ricalcolati come vettori del valore medio degli attributi delle istanze assegnate a determinati cluster.

Opzioni dei campi K-Means di Netezza

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.

Figura 6-9
Opzioni dei campi K-Means



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 di origine a monte. Per ulteriori informazioni, vedere l'argomento Impostazione del ruolo del campo in il capitolo 4 in *IBM SPSS Modeler 15 Nodi di input, elaborazione e output*.

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.

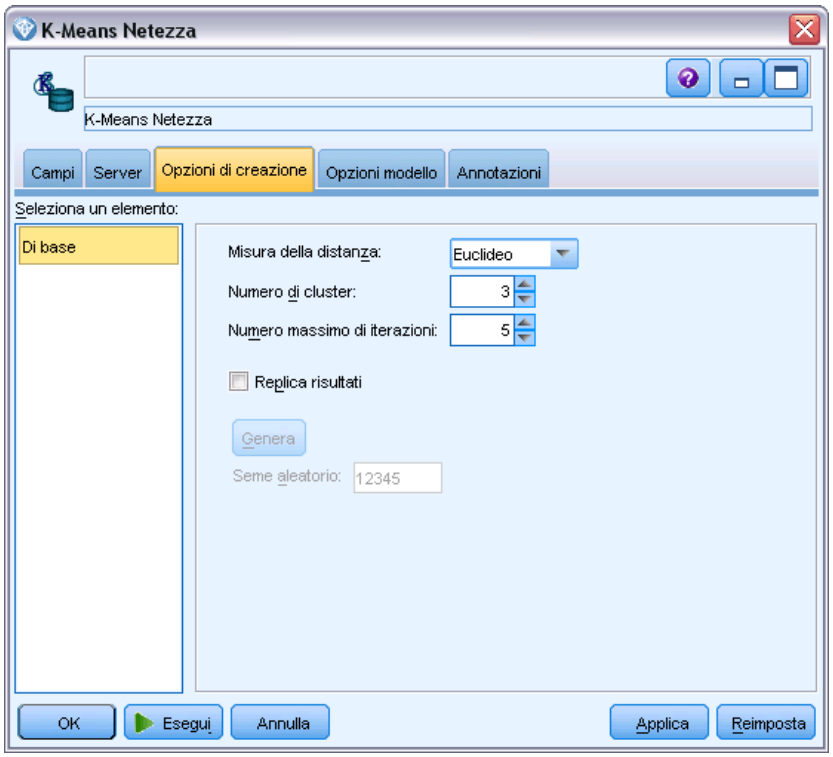
ID record. Campo da utilizzare come identificatore univoco del record.

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

Opzioni di creazione K-Means di Netezza

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.

Figura 6-10
opzioni di creazione K-Means



Misura della distanza. Metodo utilizzato per misurare la distanza tra i punti dei dati; una maggiore distanza indica una maggiore dissimilarità. Le opzioni disponibili sono:

- **Euclidea**. (default) La distanza fra due punti calcolata unendoli con una linea retta.
- **Manhattan**. La distanza fra due punti calcolata come somma delle differenze assolute fra le loro coordinate.
- **Canberra**. Simile alla distanza di Manhattan, ma più sensibile ai punti dei dati più vicini all'origine.
- **Massimo**. La distanza fra due punti calcolata come differenza massima sulla dimensione di qualsiasi coordinata.

Numero di cluster (k). Specificare il numero di cluster da creare.

Numero massimo di iterazioni. L'algoritmo funziona eseguendo una serie di iterazioni dello stesso processo. Questa opzione permette di interrompere l'addestramento del modello dopo il numero di iterazioni specificato.

Replica risultati. Selezionare questa casella per impostare un seme aleatorio che consentirà di replicare le analisi. È possibile specificare un valore intero o fare clic su Genera per creare un intero pseudocasuale.

Rete di Bayes Netezza

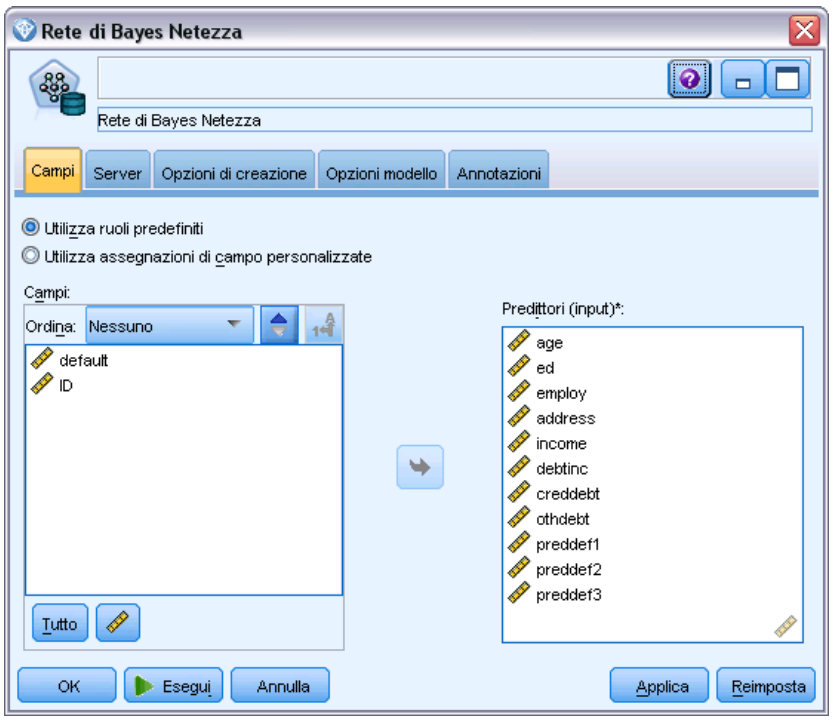
La rete bayesiana è un modello in cui sono visualizzate le variabili presenti in un insieme di dati e le indipendenze probabilistiche o condizionali tra di esse. Il nodo Rete di Bayes Netezza consente di generare un modello di probabilità combinando elementi osservati e registrati con conoscenze del mondo reale basate sul "buon senso" per stabilire la probabilità delle occorrenze utilizzando attributi apparentemente non collegati fra loro.

Opzioni dei campi della rete di Bayes Netezza

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.

Per questo nodo, il campo obiettivo è necessario solo per il calcolo del punteggio, quindi non viene visualizzato in questa scheda. È possibile impostare o modificare l'obiettivo in un nodo Tipo, nella scheda Opzioni modello di tale nodo o nella scheda Impostazioni dell'insieme di modelli. Per ulteriori informazioni, vedere l'argomento Insieme di modelli di rete di Bayes Netezza - Scheda Impostazioni a pag. 221.

Figura 6-11
Rete di Bayes, opzioni 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 di origine a monte. Per ulteriori informazioni, vedere l'argomento Impostazione del ruolo del campo in il capitolo 4 in *IBM SPSS Modeler 15 Nodi di input, elaborazione e output*.

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.

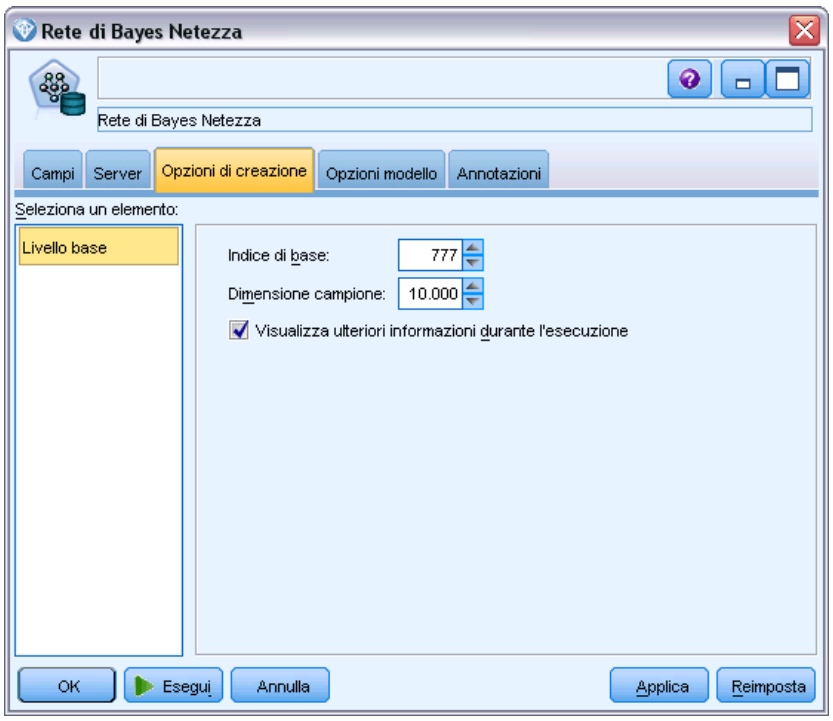
Fare clic sul pulsante Tutto per selezionare tutti i campi dell'elenco o fare clic su un singolo pulsante di livello di misurazione per selezionare tutti i campi con tale livello.

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

Opzioni di creazione della rete di Bayes Netezza

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.

Figura 6-12
Rete di Bayes, opzioni creazione



Indice di base. Identificatore numerico da assegnare al primo attributo (campo input) per semplificare la gestione interna.

Dimensione campione. Dimensione del campione da utilizzare se il numero degli attributi è talmente grande da allungare il tempo di elaborazione in modo inaccettabile.

Visualizza ulteriori informazioni durante l'esecuzione. Se questa casella è selezionata (default), vengono visualizzate ulteriori informazioni in una finestra di dialogo di messaggio.

Bayes naive Netezza

Bayes naive è un algoritmo molto noto per problemi di classificazione. Il modello viene definito *naïve* perché considera tutte le variabili di previsione proposte come indipendenti l'una dall'altra. Bayes naive è un algoritmo veloce e scalabile in grado di calcolare probabilità condizionali per combinazioni di attributi e per l'attributo obiettivo. Dai dati di addestramento viene stabilita una probabilità indipendente che serve a indicare la verosimiglianza di ciascuna classe obiettivo una volta specificata l'occorrenza di ogni categoria di valore da ogni variabile di input.

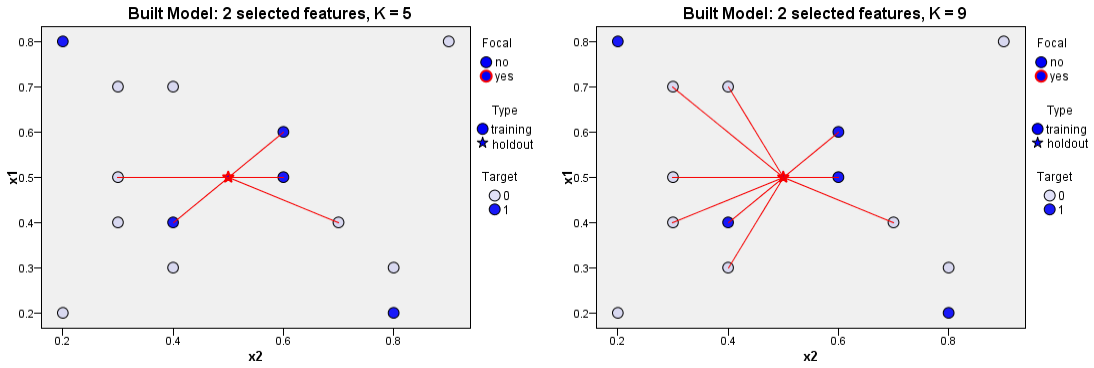
KNN Netezza

L'analisi del vicino più vicino 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 non simili sono distanti gli uni dagli altri. Pertanto, la distanza tra due casi è una misura della loro dissimilarità.

I casi che sono vicini gli uni agli altri vengono definiti "vicini". Quando viene presentato un nuovo caso (controllo), viene calcolata la sua distanza da ognuno dei casi nel modello. Le classificazioni dei casi più simili, i –vicini più vicini–, vengono contate e il nuovo caso viene posto nella categoria che contiene il maggior numero di vicini più vicini.

È possibile specificare il numero dei vicini più vicini da esaminare; questo valore viene denominato *k*. Le figure mostrano in che modo verrebbe classificato un nuovo caso utilizzando due valori diversi di *k*. Quando $k = 5$, il nuovo caso viene posto nella categoria *1* in quanto la maggior parte dei vicini più vicini appartiene alla categoria *1*. Tuttavia, quando $k = 9$, il nuovo caso viene posto nella categoria *0* in quanto la maggior parte dei vicini più vicini appartiene alla categoria *0*.

Figura 6-13
Effetti delle modifiche del valore k sulla classificazione

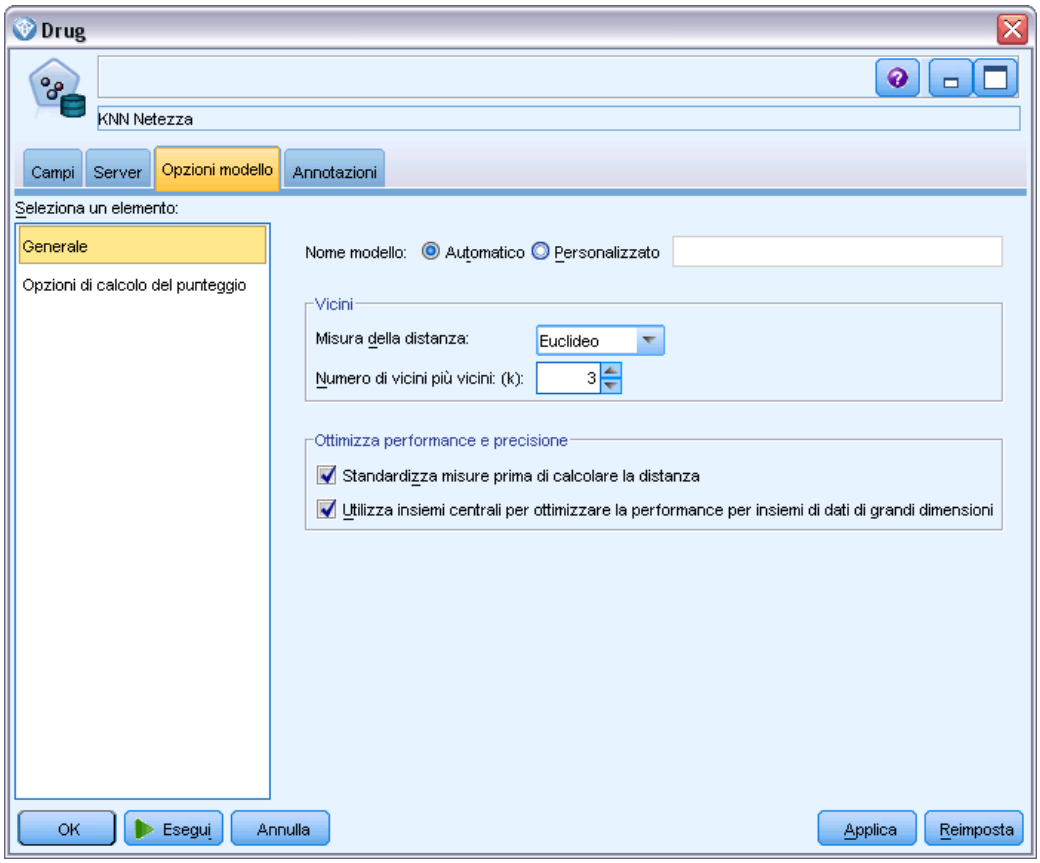


L'analisi del vicino più vicino può anche essere usata per calcolare i valori per un obiettivo continuo. In questa situazione, per ottenere il valore previsto per il nuovo caso, viene utilizzato il valore obiettivo medio o mediano dei vicini più vicini.

Opzioni del modello KNN Netezza - Generale

Nella scheda Opzioni modello - Generale è possibile scegliere se specificare un nome per il modello o generare un nome automaticamente. È inoltre possibile impostare opzioni che controllano il modo in cui il numero di vicini più vicini viene calcolato e impostare opzioni per ottimizzare la performance e la precisione del modello.

Figura 6-14
opzioni modello KNN, generali



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.

Vicini

Misura della distanza. Metodo utilizzato per misurare la distanza tra i punti dei dati; una maggiore distanza indica una maggiore dissimilarità. Le opzioni disponibili sono:

- **Euclidea**. (default) La distanza fra due punti calcolata unendoli con una linea retta.
- **Manhattan**. La distanza fra due punti calcolata come somma delle differenze assolute fra le loro coordinate.
- **Canberra**. Simile alla distanza di Manhattan, ma più sensibile ai punti dei dati più vicini all'origine.
- **Massimo**. La distanza fra due punti calcolata come differenza massima sulla dimensione di qualsiasi coordinata.

Numero di Vicini più vicini (k). Il numero di vicini più vicini relativamente a un caso specifico. L'utilizzo di un numero maggiore di vicini non garantisce necessariamente un modello più preciso.

La scelta di *k* controlla la proporzione tra la prevenzione del sovradattamento (può essere importante, soprattutto per i dati "rumorosi") e la risoluzione (con previsioni diverse per istanze simili). Normalmente è necessario adattare il valore di *k* per ogni insieme di dati; i valori tipici variano da 1 a diverse decine.

Ottimizza performance e precisione

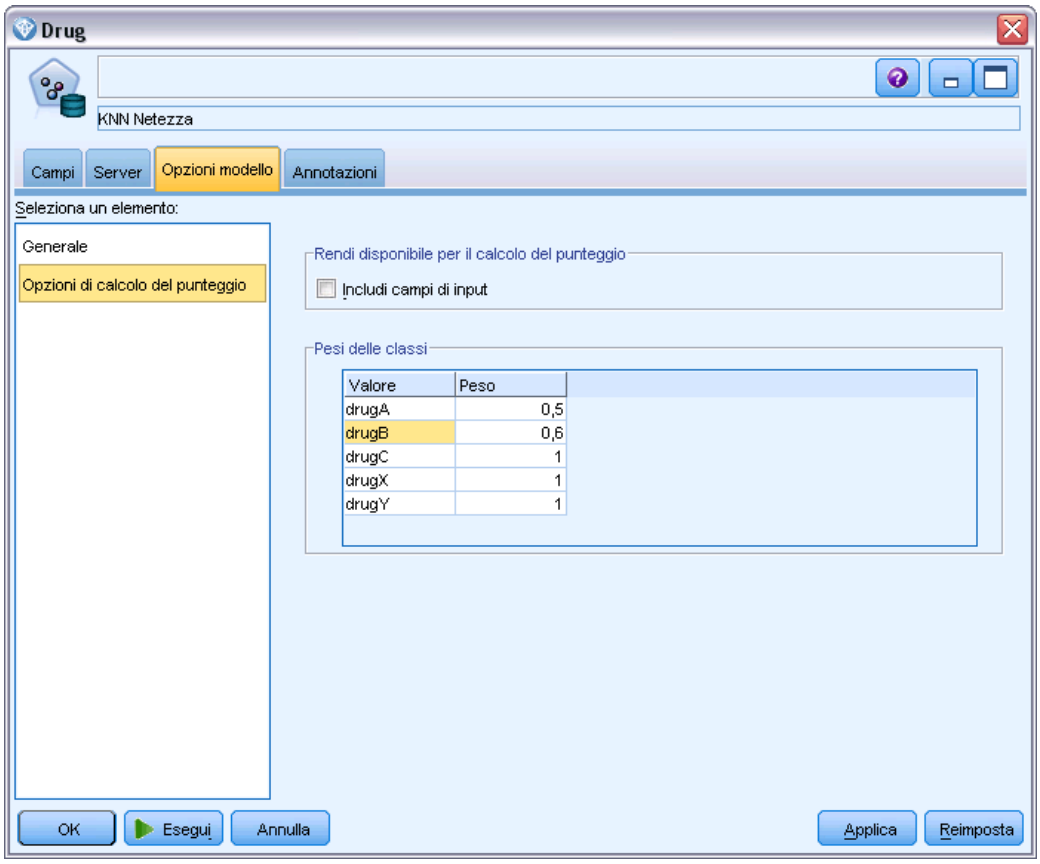
Standardizza misure prima di calcolare la distanza. Se selezionata, questa opzione standardizza le misure per i campi di input continui prima di calcolare i valori della distanza.

Utilizza insiemi centrali per ottimizzare le performance per insiemi di dati di grandi dimensioni. Se selezionata, questa opzione utilizza il campionamento degli insiemi centrali per accelerare il calcolo quando si lavora con insiemi di dati di grandi dimensioni.

Opzioni del modello KNN Netezza - Opzioni di calcolo del punteggio

Nella scheda Opzioni modello - Opzioni di calcolo del punteggio è possibile impostare il valore di default per un'opzione di calcolo del punteggio e assegnare pesi relativi alle singole classi.

Figura 6-15
opzioni modello KNN, generali



Rendi disponibile per il calcolo del punteggio

Includi campi di input. Specifica se i campi di input vengono inclusi nel calcolo del punteggio per default.

Pesi delle classi

Utilizzare questa opzione se si desidera modificare l'importanza relativa delle singole classi durante la creazione del modello.

Nota: questa opzione è abilitata solo se si utilizza il modello KNN per la classificazione. Se si esegue una regressione, ovvero il tipo di campo obiettivo è Continuo, l'opzione è disabilitata.

L'impostazione di default è assegnare il valore 1 a tutte le classi in modo che abbiano lo stesso peso. Specificando valori numerici diversi per i pesi delle singole etichette di classe, si indica all'algoritmo di pesare gli insiemi di addestramento delle singole classi.

Per modificare un peso, farvi doppio clic sopra nella colonna Peso e apportare le modifiche desiderate.

Valore. L'insieme delle etichette di classe ricavato dai valori possibili del campo obiettivo.

Peso. Il peso da assegnare a una determinata classe. L'assegnazione di un peso superiore a una classe rende il modello più sensibile a quella classe rispetto alle altre.

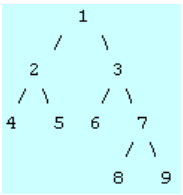
Raggruppamento cluster divisivo Netezza

Il raggruppamento cluster divisivo è un metodo di analisi in cluster in cui l'algoritmo viene eseguito ripetutamente in modo da suddividere i cluster in sottocluster finché non si raggiunge un punto di arresto specifico.

La formazione del cluster inizia con un solo cluster contenente tutte le istanze di addestramento (record). La prima iterazione dell'algoritmo divide l'insieme di dati in due sottocluster, che le successive iterazioni dividono in ulteriori sottocluster. Il criterio di arresto viene specificato come numero massimo di iterazioni, numero massimo di livelli in cui l'insieme di dati viene suddiviso e numero minimo di istanze necessarie per l'ulteriore partizionamento.

Viene generato un raggruppamento in cluster con struttura gerarchica che consente di classificare le istanze propagandole a partire dal cluster radice, come nell'esempio che segue.

Figura 6-16
Esempio di raggruppamento cluster divisivo



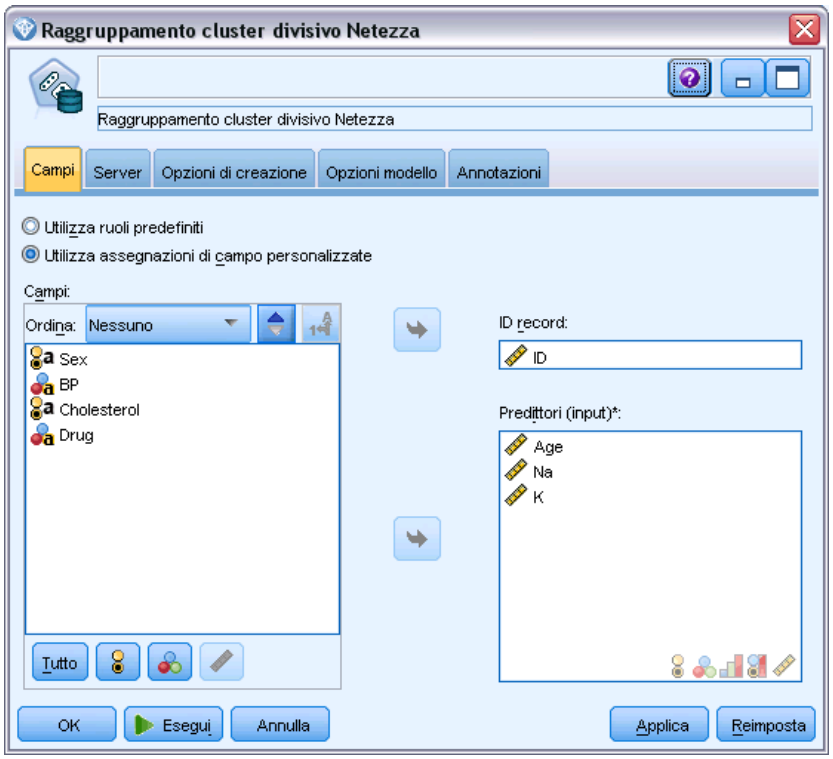
A ogni livello viene scelto il sottocluster con la migliore corrispondenza in base alla distanza dell'istanza dai centri dei sottocluster.

Quando viene calcolato il punteggio per le istanze con livello -1 della gerarchia (default), il calcolo del punteggio restituisce solo un cluster foglia, poiché le foglie sono identificate da un numero negativo. Nell'esempio, può trattarsi di uno dei cluster 4, 5, 6, 8 o 9. Tuttavia, se il livello della gerarchia è impostato su 2, per esempio, il calcolo del punteggio restituirà uno dei cluster al secondo livello sotto il cluster radice, ovvero 4, 5, 6 o 7.

Opzioni dei campi di raggruppamento cluster divisivo Netezza

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.

Figura 6-17
Raggruppamento cluster divisivo, opzioni 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 di origine a monte. Per ulteriori informazioni, vedere l'argomento Impostazione del ruolo del campo in il capitolo 4 in *IBM SPSS Modeler 15 Nodi di input, elaborazione e output*.

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.

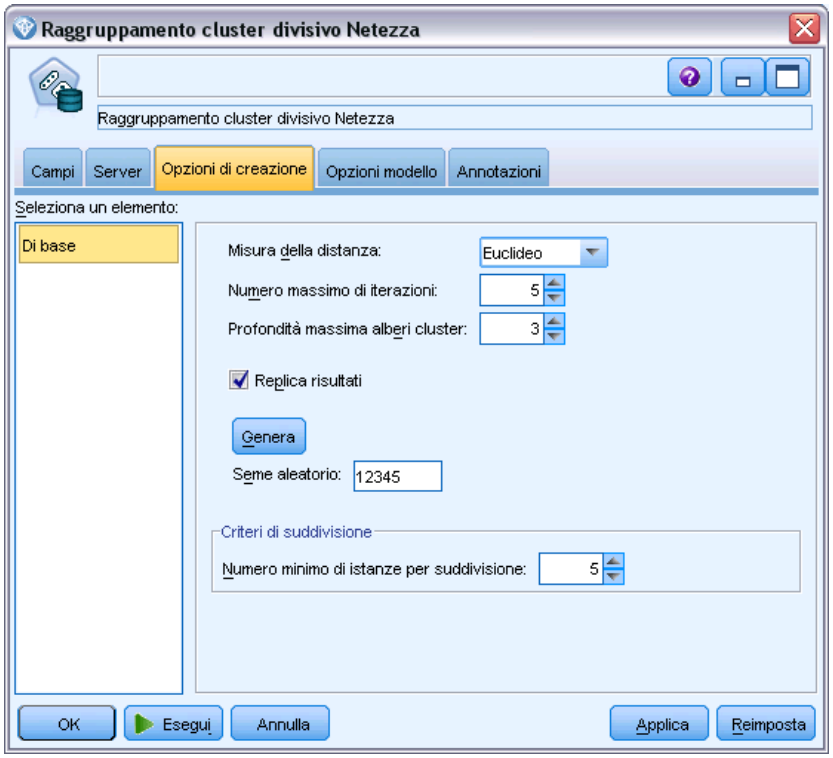
ID record. Campo da utilizzare come identificatore univoco del record.

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

Opzioni di creazione del raggruppamento cluster divisivo Netezza

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.

Figura 6-18
Raggruppamento cluster divisivo, opzioni creazione



Misura della distanza. Metodo utilizzato per misurare la distanza tra i punti dei dati; una maggiore distanza indica una maggiore dissimilarità. Le opzioni disponibili sono:

- **Euclidea**. (default) La distanza fra due punti calcolata unendoli con una linea retta.
- **Manhattan**. La distanza fra due punti calcolata come somma delle differenze assolute fra le loro coordinate.

- **Canberra**. Simile alla distanza di Manhattan, ma più sensibile ai punti dei dati più vicini all'origine.
- **Massimo**. La distanza fra due punti calcolata come differenza massima sulla dimensione di qualsiasi coordinata.

Numero massimo di iterazioni. L'algoritmo funziona eseguendo una serie di iterazioni dello stesso processo. Questa opzione permette di interrompere l'addestramento del modello dopo il numero di iterazioni specificato.

Profondità massima alberi cluster. Numero massimo di livelli in cui è possibile suddividere l'insieme di dati.

Replica risultati. Selezionare questa casella per impostare un seme aleatorio che consentirà di replicare le analisi. È possibile specificare un valore intero o fare clic su Genera per creare un intero pseudocasuale.

Numero minimo di istanze per suddivisione. Numero minimo di record che possono essere suddivisi. Quando la quantità di record ancora da suddividere è inferiore a questo numero, non verranno eseguite altre suddivisioni. È possibile utilizzare questo campo per impedire che vengano creati sottogruppi molto piccoli nell'albero dei cluster.

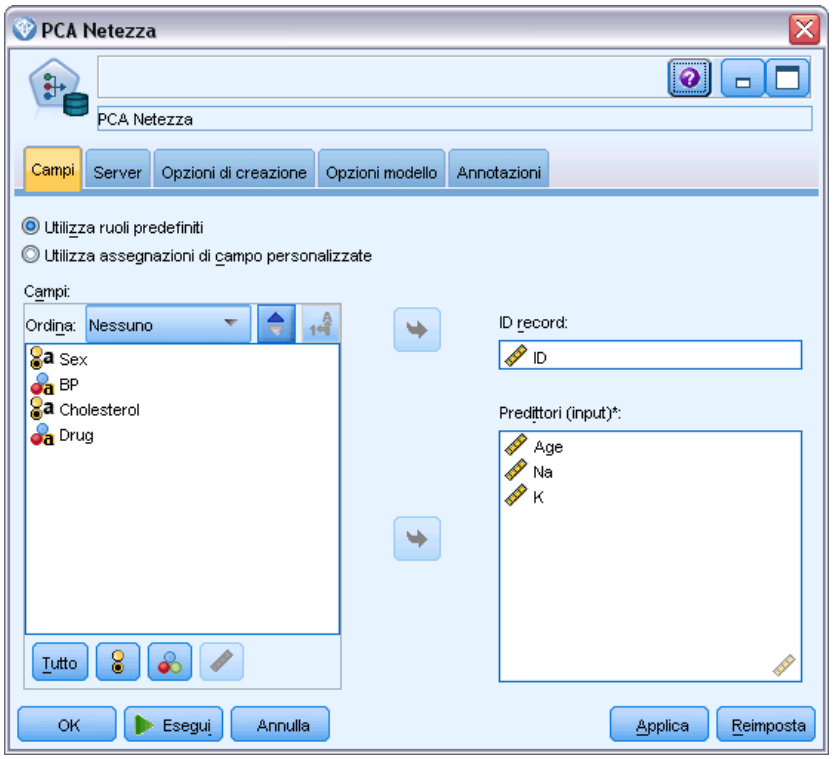
PCA Netezza

PCA (Principal Component Analysis), o analisi delle componenti principali, è una tecnica efficace progettata appositamente per ridurre la complessità dei dati. PCA trova le combinazioni lineari dei campi di input che catturano meglio la varianza nell'intero insieme di campi, dove le componenti sono ortogonali (e non correlate) le une rispetto alle altre. Lo scopo è individuare un numero limitato di campi derivati (le componenti principali) contenenti un riepilogo efficace delle informazioni presenti nell'insieme originale di campi di input.

Opzioni dei campi PCA Netezza

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.

Figura 6-19
PCA, opzioni 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 di origine a monte. Per ulteriori informazioni, vedere l'argomento Impostazione del ruolo del campo in il capitolo 4 in *IBM SPSS Modeler 15 Nodi di input, elaborazione e output*.

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.

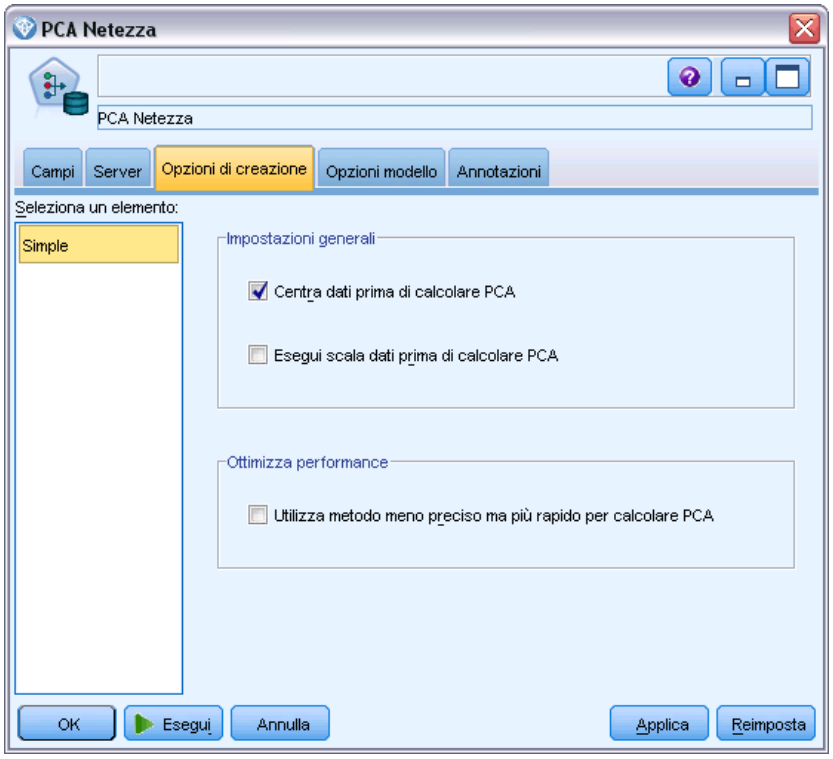
ID record. Campo da utilizzare come identificatore univoco del record.

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

Opzioni di creazione PCA Netezza

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.

Figura 6-20
PCA, opzioni di creazione



Centra dati prima di calcolare PCA. Se selezionata (default), questa opzione esegue la centratura dei dati (nota anche come "sottrazione delle medie") prima dell'analisi. La centratura dei dati è necessaria per assicurarsi che la prima componente principale descriva la direzione della varianza massima, altrimenti la componente potrebbe corrispondere maggiormente alla media dei dati. Deselezionare questa opzione per migliorare la performance solo se i dati sono già stati preparati in questo modo.

Esegui scala dati prima di calcolare PCA. Questa opzione esegue la scala dei dati prima dell'analisi. Questa operazione può ridurre l'arbitrarietà dell'analisi quando vengono misurate diverse variabili in diverse unità. La forma più semplice di scala dei dati consiste nel dividere ogni variabile per la sua variazione standard.

Utilizza metodo meno preciso ma più rapido per calcolare PCA. Questa opzione indica all'algoritmo di utilizzare un metodo meno preciso ma più rapido (forceEigensolve) per trovare le componenti principali.

Albero di regressione Netezza

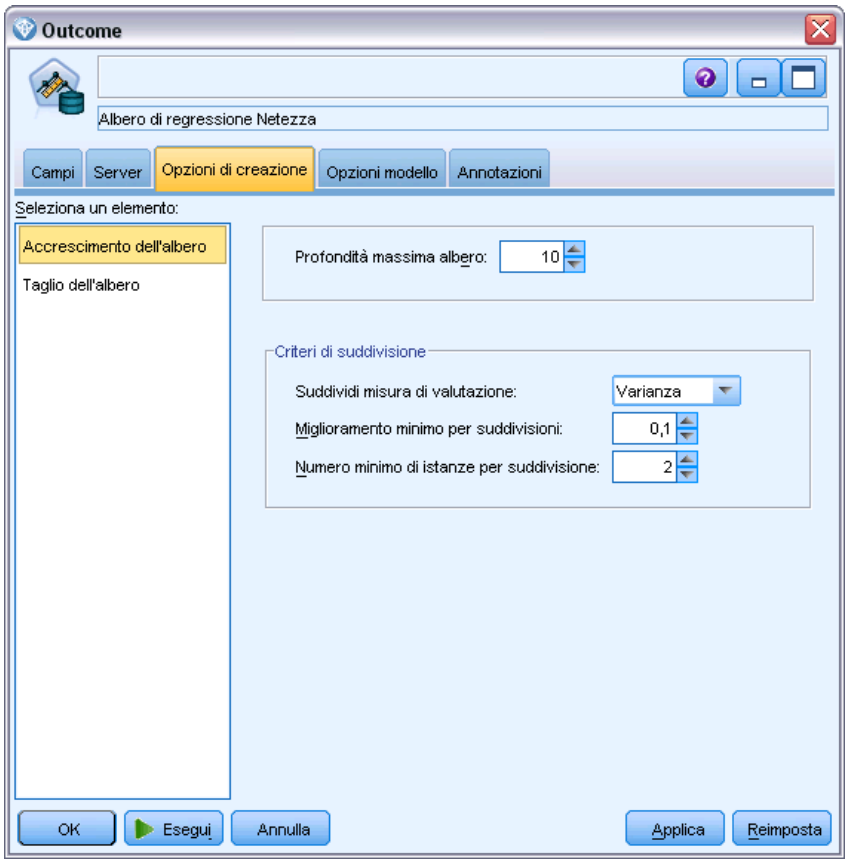
L'albero di regressione è un algoritmo basato su alberi che suddivide più volte un campione di casi per derivare sottoinsiemi dello stesso tipo, in base ai valori di un campo obiettivo numerico. Come gli alberi decisionali, gli alberi di regressione decompongono i dati in sottoinsiemi in cui le foglie dell'albero corrispondono a sottoinsiemi sufficientemente piccoli o sufficientemente uniformi. Le suddivisioni vengono selezionate in modo da ridurre la dispersione dei valori dell'attributo obiettivo e quindi consentire una previsione soddisfacente da parte dei valori medi in corrispondenza delle foglie.

L'output dei modelli assume la forma di una rappresentazione di testo dell'albero. Ogni riga di testo corrisponde a un nodo o una foglia e il rientro riflette il livello dell'albero. Per un nodo, viene visualizzata la condizione di suddivisione, per una foglia appare l'etichetta di classe assegnata.

Opzioni di creazione dell'albero di regressione Netezza - Espansione dell'albero

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.

Figura 6-21
Opzioni di creazione dell'albero di regressione per l'espansione dell'albero



Profondità massima albero. Numero massimo di foglie fino al quale l'albero può crescere sotto il nodo principale, ovvero il numero di volte che il campione può essere suddiviso in modo ricorsivo. Il valore di default è 62, che è la massima profondità possibile dell'albero ai fini della modellazione. Si noti tuttavia che il visualizzatore nell'insieme di modelli è in grado di visualizzare al massimo 12 foglie.

Criteri di suddivisione. Queste opzioni controllano il momento in cui interrompere la suddivisione dell'albero. Se non si desidera utilizzare i valori di default, fare clic su Personalizza e apportare le modifiche.

- **Suddividi misura di valutazione**. Misurazione dell'impurità della classe utilizzata per valutare il punto migliore in cui suddividere l'albero. *Nota*: attualmente Varianza è l'unica opzione possibile.
- **Miglioramento minimo per suddivisioni**. La quantità minima in base alla quale l'impurità deve essere ridotta prima di creare una nuova suddivisione nell'albero. La creazione dell'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

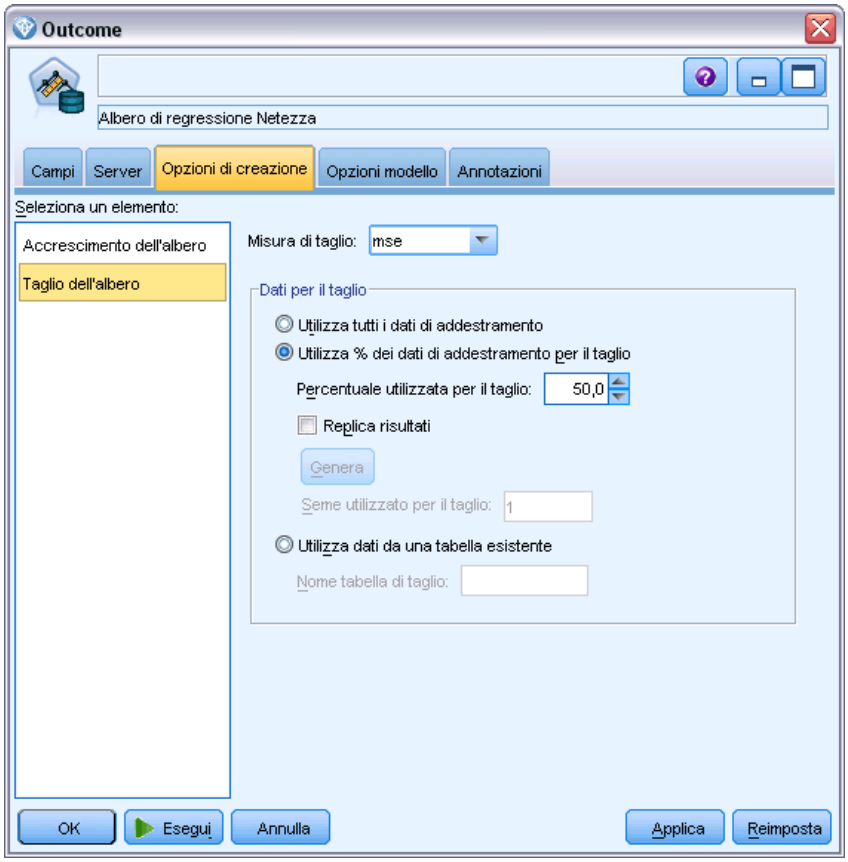
riduce l'impurità di un valore inferiore a quello specificato dal criterio di suddivisione, la suddivisione non verrà eseguita.

- **Numero minimo di istanze per suddivisione**. Numero minimo di record che possono essere suddivisi. Quando la quantità di record ancora da suddividere è inferiore a questo numero, non verranno eseguite altre suddivisioni. È possibile utilizzare questo campo per impedire che vengano creati sottogruppi molto piccoli nell'albero.

Opzioni di creazione dell'albero di regressione Netezza - Taglio dell'albero

Le opzioni di taglio consentono di specificare i criteri con cui l'albero di regressione viene tagliato. Lo scopo del taglio è ridurre il rischio di sovradattamento rimuovendo i sottogruppi cresciuti troppo che non migliorano la precisione attesa nei nuovi dati.

Figura 6-22
Opzioni di creazione dell'albero di regressione per il taglio dell'albero



Misura di taglio. La misura del taglio garantisce che la precisione stimata del modello rimanga entro limiti accettabili dopo la rimozione di una foglia dall'albero. È possibile scegliere tra le misure seguenti:

- **mse.** Errore quadratico medio (default): misura la vicinanza di una retta interpolante ai punti di dati.

- **r2.** R-quadrato: misura la proporzione di variabilità della variabile dipendente spiegata dal modello di regressione.
- **Pearson.** Coefficiente di correlazione di Pearson: misura la forza della relazione tra le variabili linearmente dipendenti che sono normalmente distribuite.
- **Spearman.** Coefficiente di correlazione di Spearman: rileva le relazioni non lineari che risultano deboli in base alla correlazione di Pearson ma che in realtà possono essere forti.

Dati per il taglio. È possibile utilizzare alcuni o tutti i dati di addestramento per stimare la precisione attesa dei nuovi dati. In alternativa, è possibile utilizzare un insieme di dati di taglio separato estratti da una tabella specifica.

- **Utilizza tutti i dati di addestramento**. Questa opzione (default) utilizza tutti i dati di addestramento per stimare la precisione del modello.
- **Utilizza % dei dati di addestramento per il taglio**. Utilizzare questa opzione per dividere i dati in due insiemi, uno per l'addestramento e uno per il taglio, usando la percentuale qui specificata per i dati del taglio.

  Selezionare Replica risultati se si desidera specificare un seme aleatorio per assicurarsi che i dati vengano partizionati nello stesso modo ogni volta che si esegue lo stream. È possibile specificare un valore intero nel campo Seme utilizzato per il taglio oppure fare clic su Genera per creare un intero pseudocasuale.
- **Utilizza dati da una tabella esistente**. Specificare il nome della tabella di un insieme di dati di taglio separato per la stima della precisione del modello. Questa operazione è più affidabile rispetto all'utilizzo dei dati di addestramento. Tuttavia, questa opzione può causare la rimozione di un grande sottoinsieme di dati dall'insieme di addestramento, riducendo la qualità dell'albero decisionale.

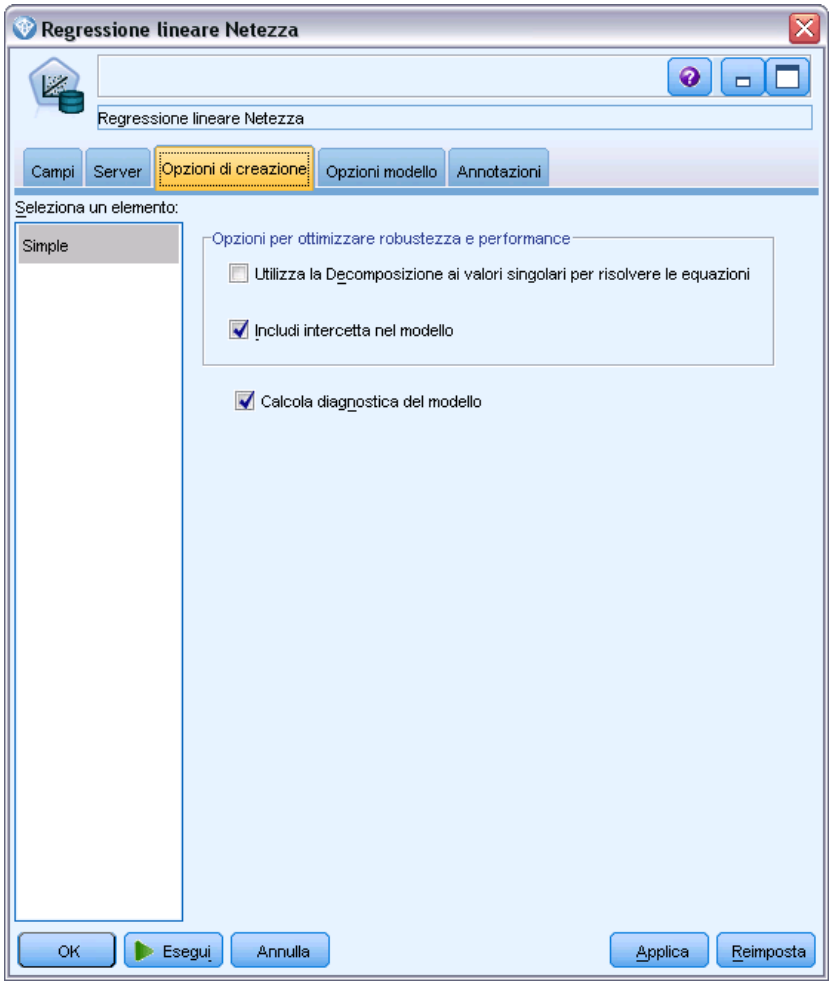
Regressione lineare Netezza

I modelli lineari prevedono un obiettivo continuo basato sulle relazioni lineari tra l'obiettivo e uno o più predittori. Benché limitati unicamente alla modellazione diretta delle relazioni lineari, i modelli di regressione lineare sono relativamente semplici e forniscono una formula matematica di facile interpretazione per il calcolo del punteggio. I modelli lineari sono veloci, efficaci e facili da utilizzare, anche se meno flessibili rispetto a quelli generati da algoritmi di regressione più sofisticati.

Opzioni di creazione della regressione lineare Netezza

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.

Figura 6-23
Opzioni di creazione della regressione lineare



Utilizza la Decomposizione ai valori singolari per risolvere le equazioni. L'utilizzo della matrice di Decomposizione ai valori singolari, anziché della matrice originale, ha il vantaggio di essere robusta rispetto agli errori numerici e allo stesso tempo velocizza il calcolo.

Includi intercetta nel modello. Includere l'intercetta aumenta la precisione generale della soluzione.

Calcola diagnostica del modello. Questa opzione consente di generare calcoli diagnostici per il modello. I risultati vengono archiviati in matrici o tabelle per la revisione successiva. La diagnostica include r-quadrato, somma dei quadrati residua, stima della varianza, deviazione standard, valore *p* e valore *t*.

La diagnostica è correlata alla validità e all'utilità del modello. È necessario eseguire la diagnostica separatamente sui dati sottostanti per garantire che soddisfi i presupposti di linearità.

Serie storica Netezza

Una **serie storica** è una sequenza di valori numerici misurati in momenti temporali successivi (anche se non necessariamente a intervalli regolari), ad esempio i prezzi giornalieri delle azioni o i dati di vendita settimanali. L'analisi di questo tipo di dati può essere utile, ad esempio, per evidenziare comportamenti che rivelano trend o stagionalità (pattern ripetuti), e per predire comportamenti futuri basandosi su eventi passati.

Serie storica Netezza supporta i seguenti algoritmi per serie storiche.

- analisi spettrale
- livellamento esponenziale
- Modello Autoregressivo Integrato a Media Mobile (ARIMA)
- scomposizione trend stagionale

Lo scopo di questi algoritmi è di estrapolare dalla serie storica un trend o un componente stagionale. I componenti verranno poi analizzati nell'ottica di creare un modello predittivo.

L'**analisi spettrale** individua i comportamenti periodici nelle serie storiche. Per le serie storiche composte di varie periodicità implicite o quando i dati contengono una quantità notevole di rumore casuale, l'analisi spettrale rappresenta il modo più chiaro per individuare i componenti periodici. Per rilevare la frequenza dei comportamenti periodici, questo metodo trasforma la serie storica dall'ambito temporale all'ambito della frequenza.

Il **livellamento esponenziale** è un metodo di previsione che utilizza i valori ponderati delle osservazioni di serie precedenti per prevedere i valori futuri. Con il livellamento esponenziale, l'influenza delle osservazioni diminuisce nel tempo in modo esponenziale. Questo metodo esegue la previsione di un punto di tempo per volta, rettificando la previsione non appena riceve nuovi dati quali aggiunte, trend e stagionalità.

I modelli **ARIMA** forniscono metodi più sofisticati per la modellazione di trend e componenti stagionali rispetto ai modelli di livellamento esponenziale Questo metodo comporta l'indicazione esplicita di ordini autoregressivi e di media mobile, nonché del grado di differenziazione.

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 storica senza cercare di spiegare le ragioni di tale comportamento.

La **scomposizione trend stagionale** elimina il comportamento periodico dalla serie storica per eseguire l'analisi del trend e quindi seleziona una forma semplice per il trend, come una funzione quadratica. Le forme semplici presentano un numero di parametri i cui valori sono determinati in modo da ridurre al minimo l'errore quadratico medio dei residui (vale a dire, le differenze tra i valori previsti e i valori osservati della serie storica).

Interpolazione dei valori nella serie storica Netezza

L'**interpolazione** è il processo di stima e inserimento dei valori mancanti nei dati di una serie storica.

Se gli intervalli di una serie storica sono regolari ma alcuni valori sono assenti, i valori mancanti potranno essere stimati mediante l'interpolazione lineare. Consideriamo la seguente serie che contiene gli arrivi mensili di passeggeri al terminal di un aeroporto.

Tabella 6-1
Arrivi mensili presso un terminal passeggeri

Mese	Passeggeri
3	3,500,000
4	3,900,000
5	-
6	3,400,000
7	4,500,000
8	3,900,000
9	5,800,000
10	6,000,000

In questo caso, l'interpolazione lineare stimerebbe il valore mancante per il mese 5 come 3.650.000 (il punto intermedio tra i mesi 4 e 6).

Gli intervalli irregolari vengono gestiti in modo diverso. Consideriamo la seguente serie di letture della temperatura.

Tabella 6-2
Letture temperatura

Data	Ora	Temperatura
2011-07-24	7:00	57
2011-07-24	14:00	75
2011-07-24	21:00	72
2011-07-25	7:15	59
2011-07-25	14:00	77
2011-07-25	20:55	74
2011-07-27	7:00	60
2011-07-27	14:00	78
2011-07-27	22:00	74

In questo caso abbiamo letture rilevate in tre momenti nel corso di tre giorni, ma in orari diversi che si ripetono solo in alcuni giorni. Inoltre, solo due giorni su tre sono consecutivi.

Questa situazione può essere gestita in due diversi modi: calcolando degli aggregati o determinando una dimensione di passo.

Gli aggregati potrebbero essere aggregati quotidiani calcolati con una formula basata sulla conoscenza semantica dei dati. Questa soluzione restituirebbe il seguente insieme di dati.

Tabella 6-3
Letture temperatura (aggregate)

Data	Ora	Temperatura
2011-07-24	24:00	69
2011-07-25	24:00	71
2011-07-26	24:00	null
2011-07-27	24:00	72

In alternativa, l'algoritmo può trattare la serie come una serie distinta e determinare un'adeguata dimensione di passo. In questo caso, la dimensione di passo determinata dall'algoritmo potrebbe essere 8 ore, il che restituirebbe quanto segue.

Tabella 6-4
Letture temperatura con dimensione di passo

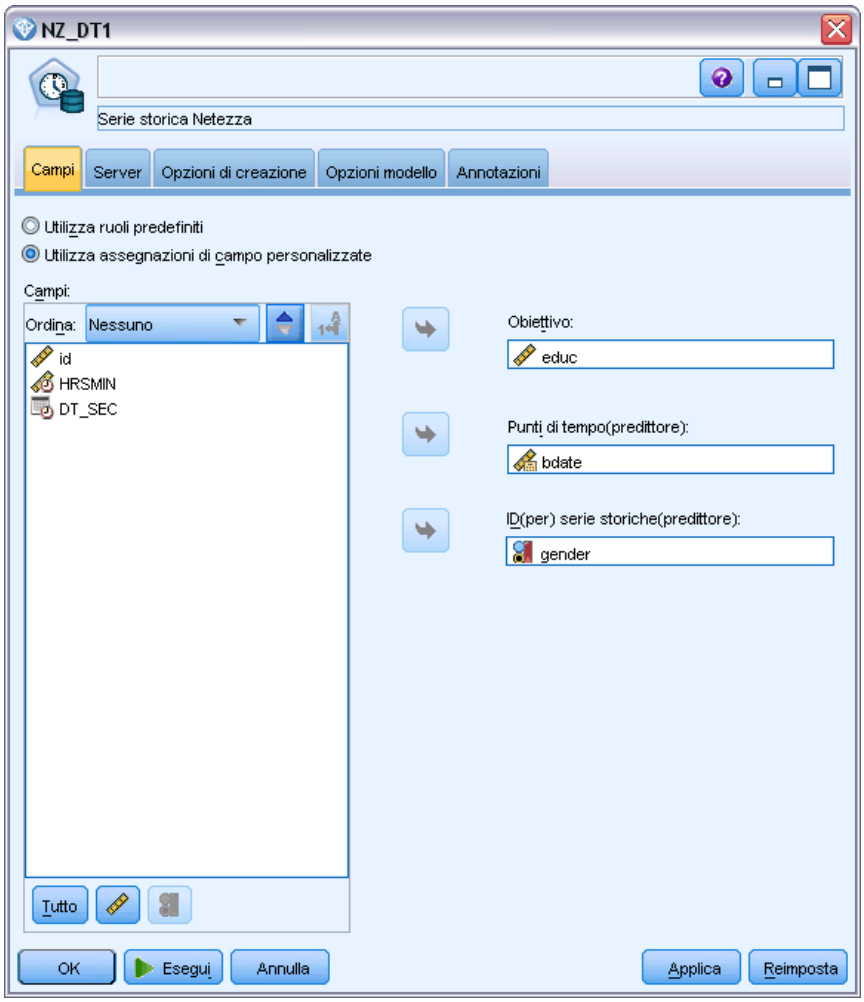
Data	Ora	Temperatura
2011-07-24	6:00	
2011-07-24	14:00	75
2011-07-24	22:00	
2011-07-25	6:00	
2011-07-25	14:00	77
2011-07-25	22:00	
2011-07-26	6:00	
2011-07-26	14:00	
2011-07-26	22:00	
2011-07-27	6:00	
2011-07-27	14:00	78
2011-07-27	22:00	74

In questo caso, solo quattro letture corrispondono alle letture originali, ma aiutandosi con altri valori conosciuti della serie originale, i valori mancanti possono essere nuovamente calcolati mediante l'interpolazione.

Opzioni dei campi della serie storica Netezza

Nella scheda Campi si specificano i ruoli per i campi di input nella sorgente dati.

Figura 6-24
Opzioni dei campi della serie storica



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. Per ulteriori informazioni, vedere l'argomento Livelli di misurazione in il capitolo 4 in *IBM SPSS Modeler 15 Nodi di input, elaborazione e output*.

Obiettivo. Scegliere un campo come obiettivo per la previsione. Questo campo deve avere un livello di misurazione Continuo.

Punti di tempo(predittore). (obbligatorio) Il campo di input contenente i valori di data o ora per la serie storica. Questo campo deve avere un livello di misurazione Continuo o Categoriale e un tipo di archiviazione dati Data, Ora, Timestamp o Numerico. Il tipo di archiviazione dati del campo specificato qui definisce anche il tipo di input di alcuni campi di altre schede di questo stesso nodo Modelli. Per ulteriori informazioni, vedere l'argomento Impostazione dell'archiviazione e della formattazione dei campi in il capitolo 2 in *IBM SPSS Modeler 15 Nodi di input, elaborazione e output*.

ID(per) serie storiche(predittore). Campo contenente ID di serie storiche; utilizzarlo se l'input contiene più di una serie storica.

Opzioni di creazione della serie storica Netezza

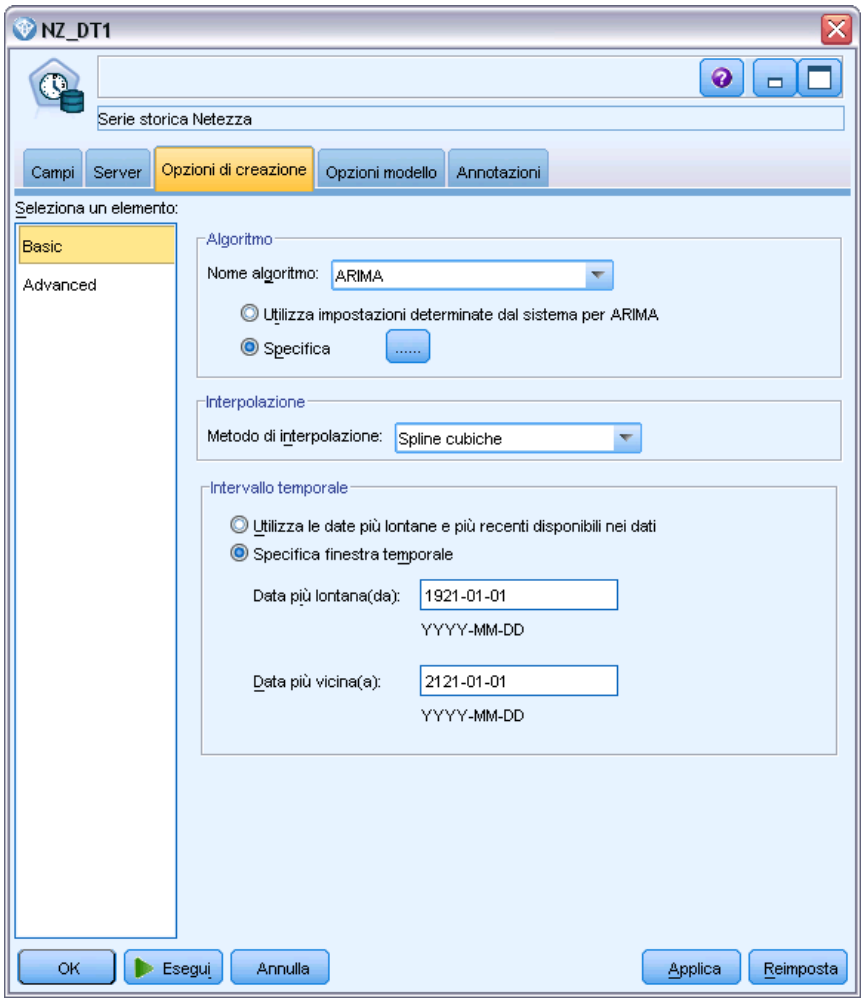
Esistono due livelli di opzioni di creazione:

- Di base - impostazioni per la scelta dell'algoritmo, dell'interpolazione e dell'intervallo di tempo da usare.
- Opzioni avanzate - impostazione per la previsione

Questa sezione descrive le opzioni di base.

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.

Figura 6-25
Opzioni di creazione di base della serie storica



Algoritmo

Sono le impostazioni relative all'algoritmo di serie storica da utilizzare.

Nome algoritmo. Scegliere l'algoritmo di serie storica da utilizzare. Gli algoritmi disponibili sono Analisi spettrale, Livellamento esponenziale (default), ARIMA e Scomposizione trend stagionale. Per ulteriori informazioni, vedere l'argomento Serie storica Netezza a pag. 200.

Trend. (solo per Livellamento esponenziale) Il livellamento esponenziale semplice non restituisce buoni risultati se la serie storica presenta un trend. Utilizzare questo campo per specificare il trend, se presente, in modo che l'algoritmo possa tenerne conto.

- **Determinato dal sistema.** (default) Il sistema tenta di trovare il valore ottimale per questo parametro.
- **Nessuno(N).** La serie storica non presenta alcun trend.

- **Additivo(A).** Trend che cresce nel tempo in maniera costante.
- **Additivo smorzato(DA).** Trend additivo che finisce per esaurirsi.
- **Moltiplicativo(M).** Trend che cresce nel tempo, generalmente in maniera più rapida rispetto a un trend additivo costante.
- **Moltiplicativo smorzato(DM).** Trend moltiplicativo che finisce per esaurirsi.

Stagionalità. (solo per Livellamento esponenziale) Utilizzare questo campo per specificare se i dati della serie storica presentano pattern stagionali.

- **Determinato dal sistema.** (default) Il sistema tenta di trovare il valore ottimale per questo parametro.
- **Nessuno(N).** La serie storica non presenta pattern stagionali.
- **Additivo(A).** Il pattern delle fluttuazioni stagionali presenta un trend costante verso l'alto nel tempo.
- **Moltiplicativo(M).** Come la stagionalità additiva, con l'aggiunta che l'ampiezza (distanza tra i punti alto e basso) delle fluttuazioni stagionali cresce rispetto al trend verso l'alto complessivo delle fluttuazioni.

Utilizza impostazioni determinate dal sistema per ARIMA. (solo per ARIMA) Scegliere questa opzione per lasciare che il sistema determini le impostazioni per l'algoritmo ARIMA.

Specifica. (solo ARIMA) Scegliere questa opzione e fare clic sul pulsante per specificare manualmente le impostazioni per ARIMA.

Interpolazione

Se i dati sorgente della serie storica presentano dei valori mancanti, scegliere un metodo che consenta di inserire valori stimati al posto dei valori mancanti. Per ulteriori informazioni, vedere l'argomento Interpolazione dei valori nella serie storica Netezza a pag. 201.

- **Lineare.** Scegliere questo metodo se gli intervalli della serie storica sono regolari e alcuni valori semplicemente non sono dati.
- **Spline esponenziali.** Applica una curva dove i punti di dati conosciuti aumentano o diminuiscono con una frequenza elevata.
- **Spline cubiche.** Applica una curva ai punti di dati conosciuti per stimare i valori mancanti.

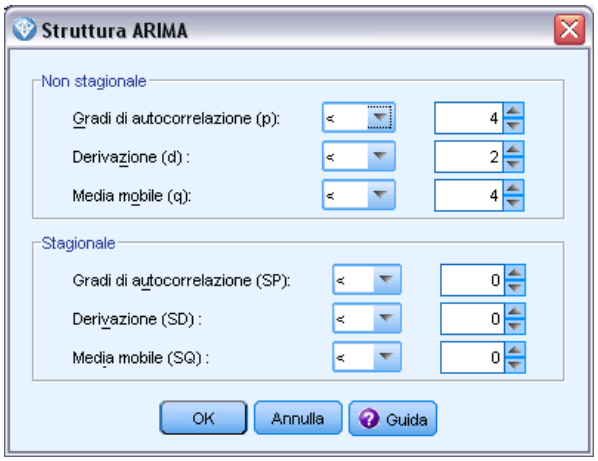
Intervallo temporale

Qui è possibile scegliere se, per creare il modello, si vuole utilizzare la gamma completa dei dati della serie storica o un sottoinsieme contiguo di tali dati. Gli input validi per questi campi sono definiti dal tipo di archiviazione dei dati specificato per i Punti di tempo della scheda Campi. Per ulteriori informazioni, vedere l'argomento Opzioni dei campi della serie storica Netezza a pag. 202.

- **Utilizza le date più lontane e più recenti disponibili nei dati.** Scegliere questa opzione per utilizzare la gamma completa dei dati della serie storica.
- **Specifica finestra temporale.** Scegliere questa opzione per utilizzare solo una porzione della serie storica. Utilizzare i campi Data più lontana(da) e Data più vicina(a) per specificare i limiti della gamma.

Struttura ARIMA

Figura 6-26
Impostazioni ARIMA per la serie storica



Specificare i valori dei vari componenti stagionali e non stagionali del modello ARIMA. In ogni caso, impostare l'operatore su < (minore di), = (uguale a) o <= (minore o uguale a), quindi immettere il valore nel campo adiacente. I valori devono specificare i gradi ed essere interi non negativi.

Non stagionale. I valori dei vari componenti non stagionali del modello.

- **Gradi di autocorrelazione (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.
- **Derivazione (d).** Specifica l'ordine di differenziazione applicato alla serie prima di eseguire la stima dei modelli. La differenziazione è necessaria quando sono presenti dei trend (di norma, le serie che presentano dei trend sono non stazionarie e il nei modelli ARIMA si presume che vi sia stazionarietà) e viene utilizzata per rimuoverne l'effetto. L'ordine di differenziazione corrisponde al grado di trend della serie, la differenziazione di primo grado tiene conto dei trend lineari, la differenziazione di secondo grado tiene conto dei trend quadratici 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.

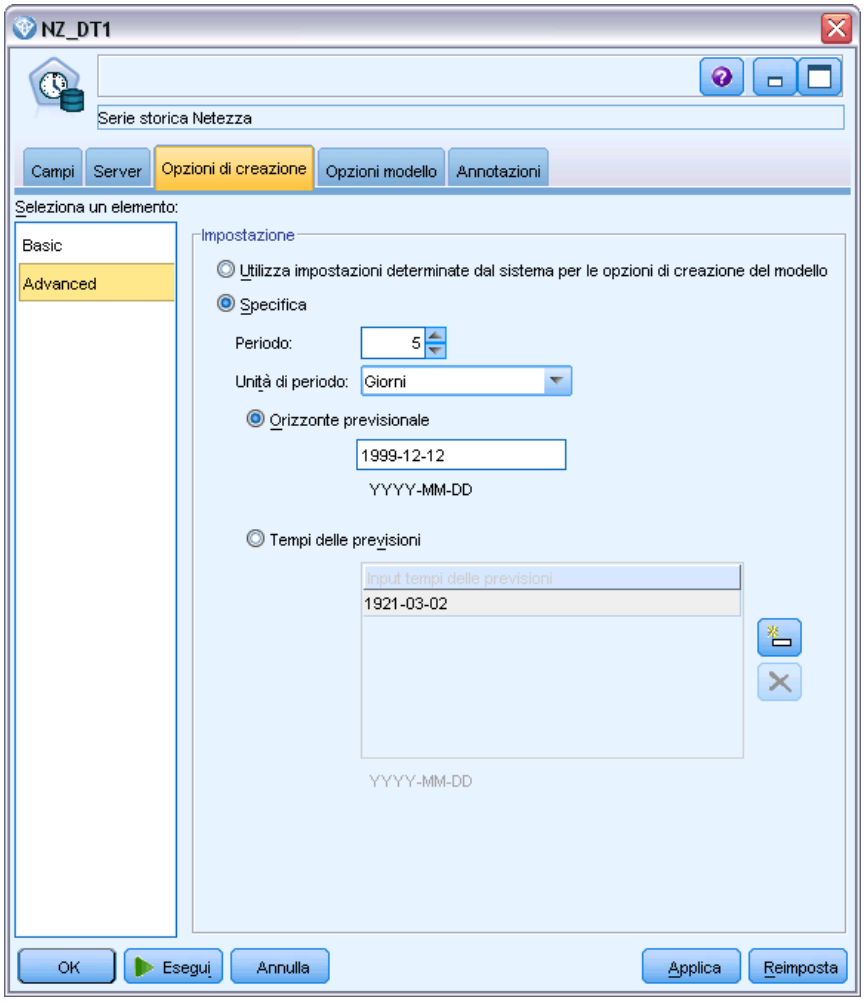
Stagionale. I componenti Autocorrelazione stagionale (SP), Derivazione (SD) e Media mobile (SQ) svolgono le stesse funzioni delle rispettive controparti non stagionali. Per gli ordini stagionali tuttavia, i valori di serie correnti vengono influenzati dai valori di serie precedenti separati da uno o 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.

Le impostazioni stagionali sono prese in considerazione solo se la stagionalità è rilevata nei dati oppure se si specificano impostazioni di Periodo nella scheda Avanzate.

Opzioni di creazione della serie storica Netezza - Avanzate

Le impostazioni avanzate consentono di specificare opzioni per la previsione.

Figura 6-27
Opzioni di creazione avanzate della serie storica



Utilizza impostazioni determinate dal sistema per le opzioni di creazione del modello. Selezionare questa opzione per lasciare che il sistema determini le impostazioni avanzate.

Specifica. Selezionare questa opzione per specificare manualmente le opzioni avanzate. (Questa opzione non è disponibile se l'algoritmo è Analisi spettrale.)

- **Periodo/Unità di periodo.** Il periodo di tempo trascorso il quale un dato comportamento caratteristico della serie storica si ripete. Ad esempio, per una serie storica di dati di vendita settimanali si definirebbe 1 per il periodo e Settimane per le unità. Periodo deve essere un intero non negativo; Unità di periodo può essere Millisecondi, Secondi, Minuti, Ore, Giorni, Settimane, Trimestri o Anni. Non impostare Unità di periodo se Periodo non è impostato oppure se il tipo di ora non è numerico. Tuttavia, se si specifica Periodo, è necessario indicare anche Unità di periodo.

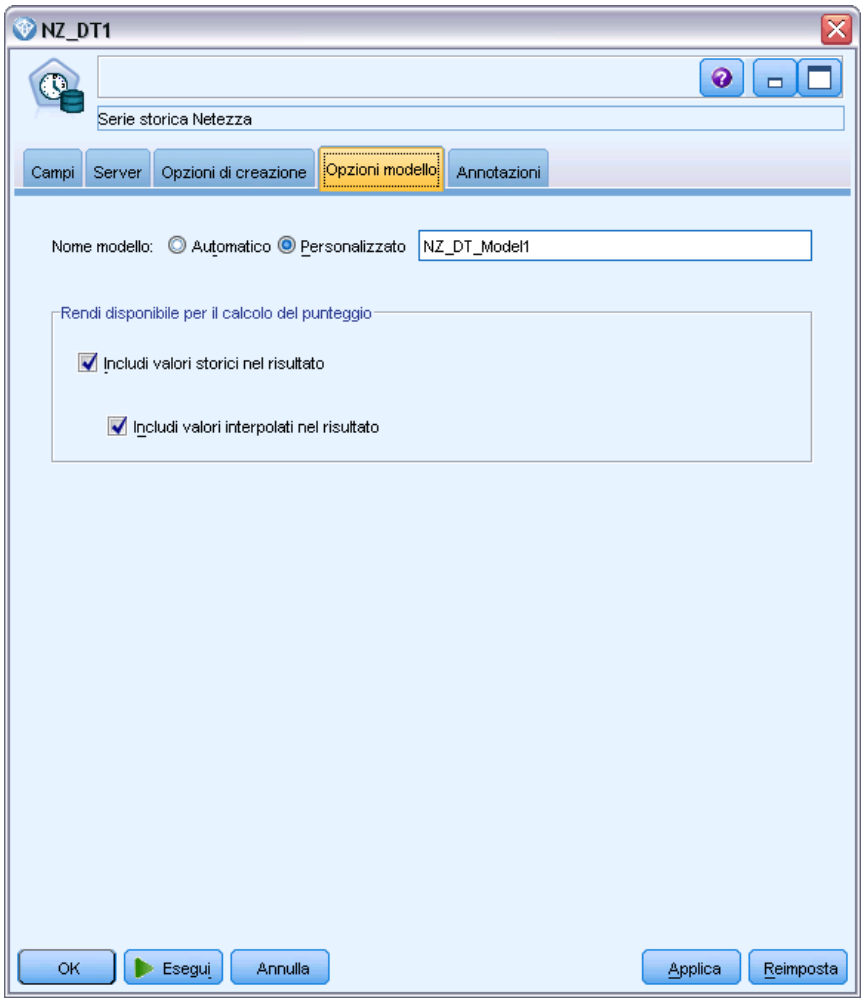
Impostazioni per la previsione. Si può scegliere di effettuare previsioni fino a uno specifico punto di tempo o in precisi punti di tempo. Gli input validi per questi campi sono definiti dal tipo di archiviazione dei dati specificato per i Punti di tempo della scheda Campi. Per ulteriori informazioni, vedere l'argomento Opzioni dei campi della serie storica Netezza a pag. 202.

- **Orizzonte previsionale.** Selezionare questa opzione per specificare solo un punto finale in cui interrompere la previsione. Le previsioni verranno effettuate fino a questo punto di tempo.
- **Tempi delle previsioni.** Selezionare questa opzione per specificare uno o più punti di tempo per cui effettuare delle previsioni. Fare clic su Aggiungi per aggiungere una nuova riga alla tabella dei punti di tempo. Per eliminare una riga, selezionarla e fare clic su Elimina.

Opzioni del modello di serie storica Netezza

Nella scheda Opzioni modello è possibile scegliere se specificare un nome per il modello o generare un nome automaticamente. È inoltre possibile impostare valori di default per le opzioni di output del modello.

Figura 6-28
Opzioni del modello di serie storica



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.

Rendi disponibile per il calcolo del punteggio. È possibile impostare qui i valori di default per le opzioni di calcolo del punteggio che appaiono nella finestra di dialogo dell'insieme di modelli.

- **Includi valori storici nel risultato.** Di default, l'output del modello non include i valori storici (quelli utilizzati per effettuare la previsione). Selezionare questa casella di controllo per includere questi valori.
- **Includi valori interpolati nel risultato.** Se si opta per includere valori storici nell'output, selezionare questa casella per includere anche i valori interpolati, se presenti. Si noti che, poiché l'interpolazione funziona solo con dati storici, questa casella non è disponibile se Includi valori storici nel risultato non è selezionata. Per ulteriori informazioni, vedere l'argomento Interpolazione dei valori nella serie storica Netezza a pag. 201.

Lineare generalizzato Netezza

La regressione lineare è una tecnica statistica impiegata da molto tempo 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. I modelli lineari sono utili per modellare un'ampia gamma di fenomeni del mondo reale in virtù della loro semplicità di addestramento e di applicazione ai modelli. Tuttavia, i modelli lineari presuppongono una distribuzione normale nella variabile dipendente (obiettivo) e un impatto lineare delle variabili indipendenti (predittori) sulla variabile dipendente.

Esistono molte situazioni in cui una regressione lineare risulta utile ma in cui i presupposti esposti sopra non sono applicabili. Ad esempio, quando si esegue la modellazione delle scelte dei consumatori tra un numero di prodotti discreto, è probabile che la variabile dipendente abbia una distribuzione multinomiale. Analogamente, quando si esegue la modellazione del reddito rispetto all'età, il reddito generalmente cresce al crescere dell'età, ma è difficile che il collegamento tra i due elementi sia una semplice linea retta.

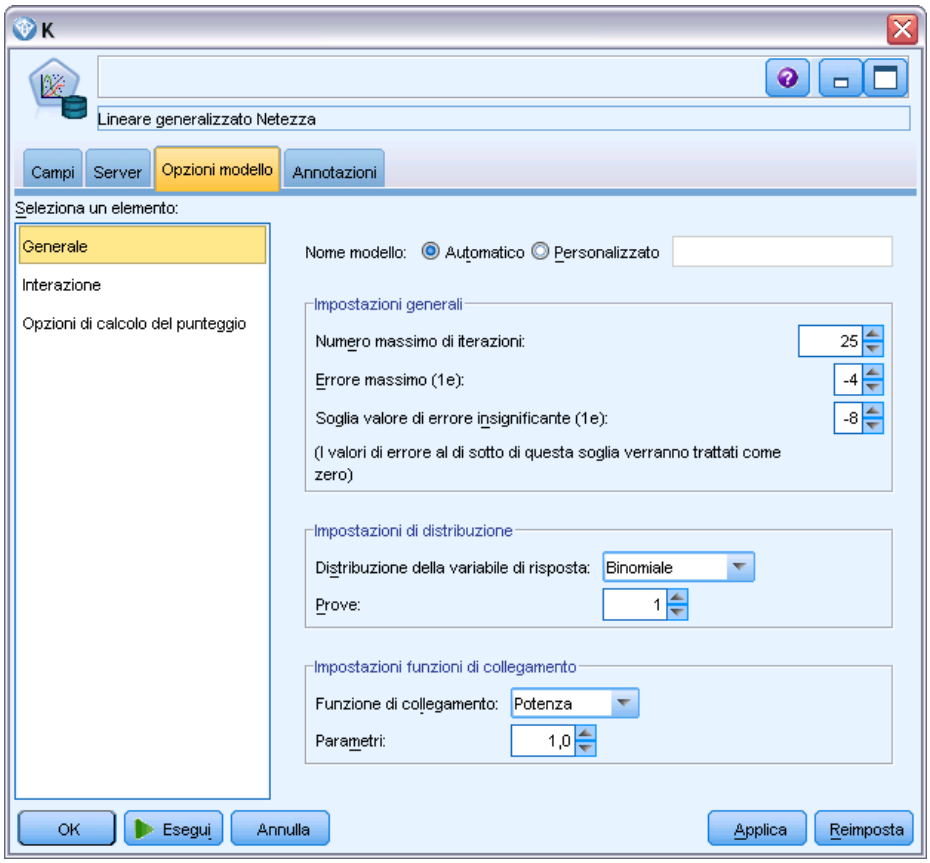
Per questi scenari è possibile utilizzare un modello lineare generalizzato. I modelli lineari generalizzati ampliano il modello di regressione lineare in modo che la variabile dipendente sia correlata alle variabili predittore per mezzo di una funzione di collegamento specifica, per cui esiste una scelta di funzioni adatte. Inoltre, il modello consente alla variabile dipendente di avere una distribuzione non normale, come la distribuzione di Poisson, binomiale e così via.

L'algoritmo esegue una ricerca iterativa del modello più adatto arrivando al numero massimo di iterazioni specificato. Per il calcolo del modello più adatto, l'errore è rappresentato dalla somma dei quadrati delle differenze tra il valore previsto e il valore attuale della variabile dipendente.

Opzioni del modello lineare generalizzato Netezza - Generale

Nella scheda Opzioni modello è possibile scegliere se specificare un nome per il modello o generare un nome automaticamente. Si possono definire anche varie impostazioni relative al modello, alla funzione di collegamento, alle interazioni tra i campi di input (se presenti) e impostare i valori di default per le opzioni di calcolo del punteggio.

Figura 6-29
Opzioni generali dei modelli lineari generalizzati



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.

Impostazioni generali. Queste impostazioni si riferiscono ai criteri di arresto dell'algoritmo.

- **Numero massimo di iterazioni.** Numero massimo di iterazioni eseguite dall'algoritmo; il minimo è 1, il default è 20.
- **Errore massimo (1e).** Il valore di errore massimo (in notazione scientifica) raggiunto il quale l'algoritmo deve interrompere la ricerca del modello più adatto. Il minimo è 0, il default è -3, vale a dire 1E-3 oppure 0,001.
- **Soglia valore di errore insignificante (1e).** Il valore (in notazione scientifica) sotto il quale gli errori vengono trattati come se avessero valore zero. Il minimo è -1, il default è -7, vale a dire che i valori inferiori a 1E-7 (o 0,0000001) sono considerati insignificanti.

Impostazioni di distribuzione. Queste impostazioni si riferiscono alla distribuzione della variabile dipendente (obiettivo)

- **Distribuzione della variabile di risposta.** Il tipo di distribuzione; uno tra Bernoulli (default), Gaussiana, Poisson, Binomiale, Binomiale negativa, Gaussiana inversa e Gamma.

- **Prove.** (Solo distribuzione Binomiale, se richiesta) Quando la risposta obiettivo è un numero di eventi che si verificano in una serie di prove, il campo obiettivo contiene il numero di eventi e il campo Prove contiene 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. Il numero di prove deve essere un numero intero positivo maggiore o uguale al numero di eventi di ciascun record.
- **Parametri.** (solo per la distribuzione binomiale negativa) È possibile specificare un valore di parametro se la distribuzione è binomiale negativa. Specificare un valore o scegliere il valore di default -1.

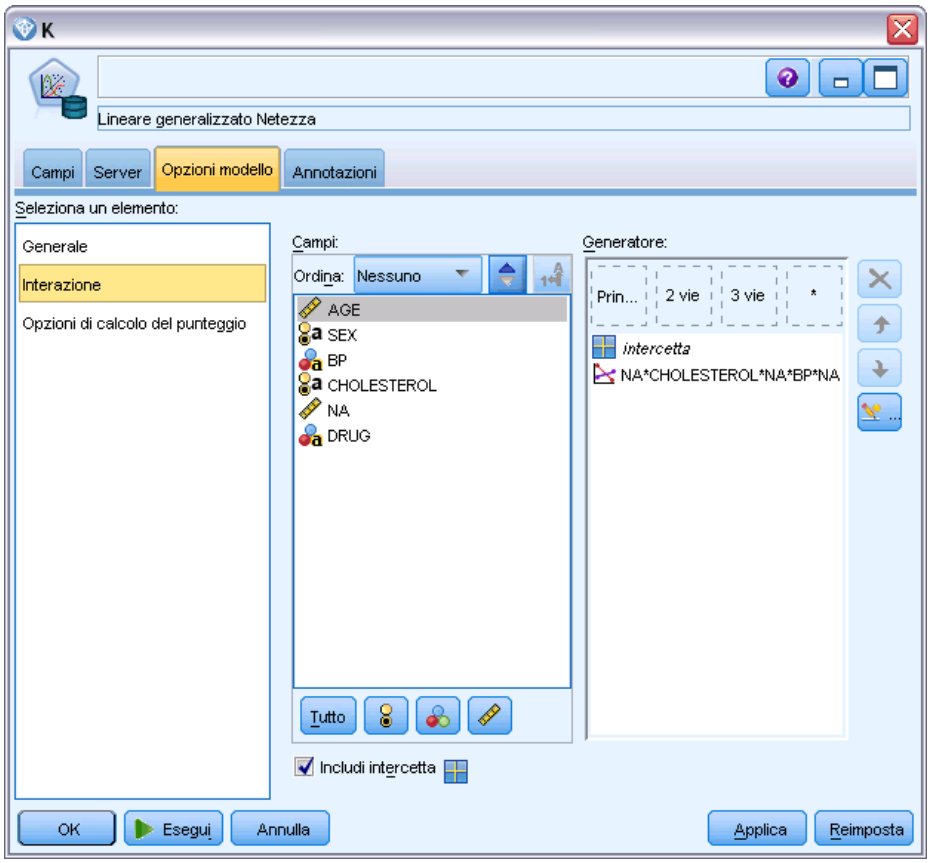
Impostazioni funzioni di collegamento. Queste impostazioni si riferiscono alla funzione di collegamento, che pone in correlazione la variabile dipendente con le variabili predittore.

- **Funzione di collegamento.** La funzione da utilizzare; una tra Identità, Inversa, Invnegative, Invsquare, Sqrt, Potenza, Oddspower, Log, Clog, Loglog, Cloglog, Logit (default), Probit, Gaussit, Cauchit, Canbinom, Cangeom, Cannegbinom.
- **Parametri.** (solo per le funzioni di collegamento Potenza e Oddspower) È possibile specificare un valore di parametro se la funzione di collegamento è Potenza o Oddspower. Specificare un valore o scegliere il valore di default 1.

Opzioni del modello lineare generalizzato Netezza - Interazione

Il riquadro Interazione contiene le opzioni per la definizione delle interazioni, vale a dire gli effetti moltiplicativi tra i campi di input.

Figura 6-30
Opzioni di interazione dei modelli lineari generalizzati



Interazione tra colonne. Selezionare questa casella di controllo per specificare le interazioni tra i campi di input. Lasciare la casella vuota se non sono presenti interazioni.

Immettere le interazioni nel modello selezionando uno o più campi nell'elenco di sorgenti e trascinandoli nell'elenco delle interazioni. Il tipo di interazione creato dipende dall'area sensibile nella quale si rilascia l'interazione.

- **Principale.** I campi rilasciati vengono visualizzati come interazioni principali separate in fondo all'elenco delle interazioni.
- **2 vie.** Tutte le possibili coppie dei campi rilasciati vengono visualizzate come interazioni a 2 vie in fondo all'elenco delle interazioni.
- **3 vie.** Tutti i possibili gruppi di tre dei campi rilasciati vengono visualizzati come interazioni a 3 vie in fondo all'elenco delle interazioni.
- ***.** La combinazione di tutti i campi rilasciati viene visualizzata come un'interazione singola in fondo all'elenco delle interazioni.

I pulsanti a destra del display consentono di:

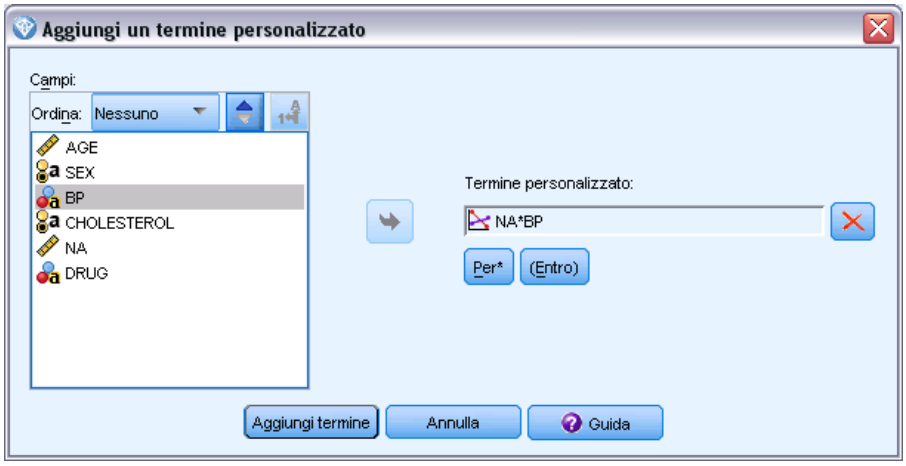
Eliminare termini dal modello selezionando i termini da eliminare e facendo clic sul pulsante di eliminazione.

Riordinare termini nel modello selezionando i termini da riordinare e facendo clic sul pulsante freccia su o giù.

Includi intercetta. L'intercetta viene in genere inclusa nel modello. Se è possibile presumere che i dati passino attraverso l'origine, l'intercetta può essere esclusa.

Aggiungi termine personalizzato

Figura 6-31
Finestra di dialogo Aggiungi termine personalizzato



Si possono specificare interazioni personalizzate nel formato *n1*x1*x1*x1...* Selezionare un campo dall'elenco Campi, fare clic sul pulsante con la freccia destra per aggiungere il campo a Termine personalizzato, fare clic su Per*, selezionare il campo successivo, fare clic sul pulsante con la freccia destra e così via. Dopo aver creato l'interazione personalizzata, fare clic su Aggiungi termine per riportarlo al riquadro Interazione.

Opzioni del modello lineare generalizzato Netezza - Opzioni di calcolo del punteggio

Rendi disponibile per il calcolo del punteggio. È possibile impostare qui i valori di default per le opzioni di calcolo del punteggio che appaiono nella finestra di dialogo dell'insieme di modelli. Per ulteriori informazioni, vedere l'argomento Insieme di modelli lineari generalizzati Netezza - Scheda Impostazioni a pag. 233.

- **Includi campi di input.** Selezionare questa casella di controllo per visualizzare i campi di input nell'output del modello oltre alle previsioni.

Gestione di modelli di IBM Netezza Analytics

I modelli IBM® Netezza® Analytics vengono aggiunti all'area di disegno e alla palette Modelli secondo modalità analoghe agli altri modelli di IBM® SPSS® Modeler e si possono utilizzare praticamente nello stesso modo. Tuttavia, esistono alcune importanti differenze, dato che ogni modello Netezza Analytics creato in SPSS Modeler fa in realtà riferimento a un modello archiviato in un server di database. Quindi affinché uno stream funzioni correttamente deve connettersi al database su cui è stato creato il modello, e la tabella del modello non deve essere stata cambiata da un processo esterno.

Calcolo del punteggio dei modelli IBM Netezza Analytics

I modelli sono rappresentati nell'area di disegno da un'icona di insieme di modelli dorata. Scopo principale degli insiemi di modelli è calcolare il punteggio dei dati per generare previsioni o consentire ulteriori analisi delle proprietà del modello. I punteggi vengono aggiunti sotto forma di uno o più campi di dati aggiuntivi che possono essere visualizzati allegando un nodo Tabella all'insieme di modelli ed eseguendo quel ramo dello stream, come descritto nella sezione che segue. Alcune finestre di dialogo degli insiemi, ad esempio quelle relative all'albero decisionale o all'albero di regressione, contengono anche una scheda Modello che fornisce una rappresentazione grafica del modello.

I campi aggiuntivi sono identificabili tramite il prefisso $<id>- aggiunto al nome del campo obiettivo, dove <id> dipende dal modello e specifica il tipo di informazioni aggiunte. I diversi identificatori vengono descritti negli argomenti per ogni insieme di modelli.

Per visualizzare i punteggi, completare la seguente procedura:

- Collegare un nodo Tabella all'insieme di modelli.
- Aprire il nodo Tabella.
- Fare clic su Esegui.
- Scorrere verso destra nella finestra di output della tabella per visualizzare i campi aggiuntivi e i relativi punteggi.

Scheda Server dell'insieme di modelli Netezza

Nella scheda Server è possibile impostare le opzioni del server per il calcolo del punteggio del modello. È possibile continuare a utilizzare una connessione al server specificata a monte oppure spostare i dati in un altro database specificato qui.

Figura 6-32
Esempio di opzioni del server dell'insieme di modelli Netezza

Dettagli Server DB Netezza. Qui si specificano i dettagli della connessione per il database da utilizzare per il modello.

- **Utilizza connessione a monte**. (default) Utilizza i dettagli di connessione specificati in un nodo a monte, per esempio il nodo di input Database. *Nota*: questa opzione funziona solo se tutti i nodi a monte sono in grado di utilizzare il push back SQL. In questo caso non è necessario spostare i dati fuori dal database perché SQL supporta pienamente tutti i nodi a monte.
- **Sposta dati alla connessione**. Sposta i dati nel database indicato qui. In questo modo si consente al modello di lavorare se i dati si trovano su un altro database IBM Netezza, o su un database di un altro fornitore, o anche se i dati si trovano in un file piatto. Inoltre i dati vengono riportati in questo database se sono stati in precedenza estratti perché un nodo non ha effettuato il push back SQL. Fare clic sul pulsante Modifica per reperire e selezionare una connessione. *Attenzione*: IBM® Netezza® Analytics è utilizzato generalmente con insiemi di dati molto grandi. Il trasferimento di grandi quantità di dati tra database, o anche dentro e fuori lo stesso database, può richiedere molto tempo ed è quindi da evitare se possibile.

Nome tabella. Nome della tabella di database in cui viene archiviato il modello. Il nome qui ha puramente scopo informativo e non può essere modificato.

Insiemi di modelli Albero decisionale di Netezza

L'insieme di modelli dell'albero decisionale visualizza l'output prodotto dall'operazione di modellazione e consente anche di impostare alcune opzioni per il calcolo del punteggio del modello.

Quando si esegue uno stream contenente un nodo Modelli dell'albero decisionale, il nodo aggiunge per default un nuovo campo, il cui nome viene derivato dal nome del modello.

Tabella 6-5
Campo di calcolo del punteggio dei modelli per l'albero decisionale

Nome del campo aggiunto	Significato
$I-*nome_modello*	Valore previsto per il record corrente.

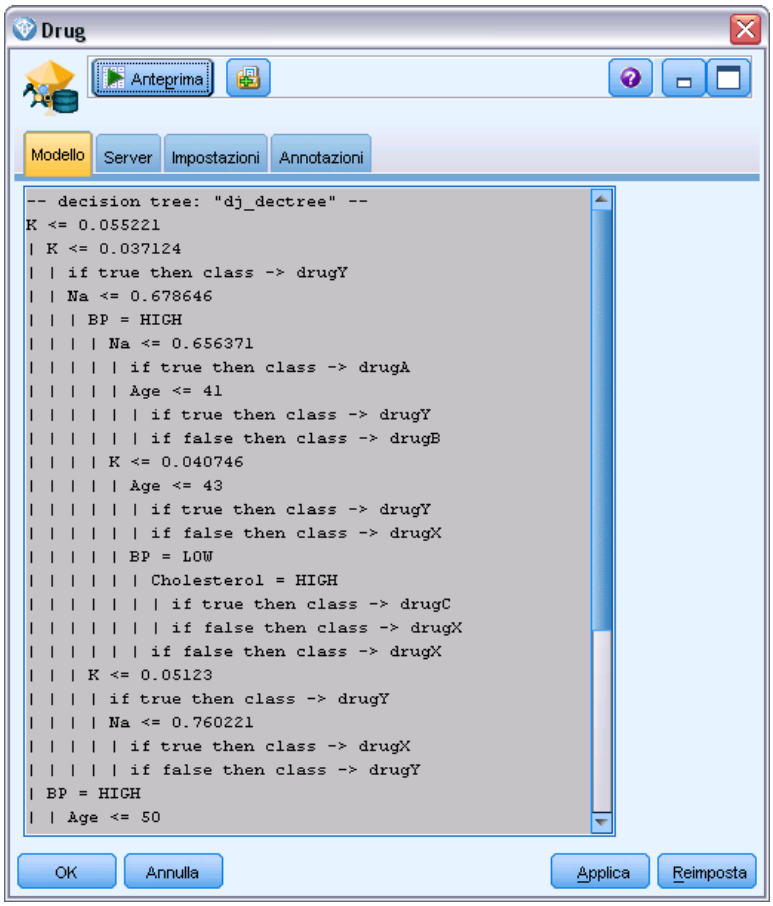
Se si seleziona l'opzione Calcola probabilità di classi assegnate per il punteggio dei record nel nodo Modelli o nell'insieme di modelli e si esegue lo stream, viene aggiunto un ulteriore campo.

Tabella 6-6
Campo di calcolo del punteggio dei modelli per l'albero decisionale - aggiuntivo

Nome del campo aggiunto	Significato
$IP-*nome_modello*	Valore di confidenza (da 0,0 a 1,0) per la previsione.

Insieme di modelli dell'albero decisionale di Netezza - Scheda Modello

Figura 6-33
Output dei modelli dell'albero decisionale



L'output dei modelli assume la forma di una rappresentazione di testo dell'albero. Ogni riga di testo corrisponde a un nodo o una foglia e il rientro riflette il livello dell'albero. Per un nodo, viene visualizzata la condizione di suddivisione, per una foglia appare l'etichetta di classe assegnata.

Insieme di modelli dell'albero decisionale di Netezza - Scheda Impostazioni

La scheda Impostazioni consente di impostare alcune opzioni di calcolo del punteggio del modello.

Includi campi di input. Se selezionata, questa opzione passa a valle tutti i campi di input originali, accodando il campo o i campi di modellazione in più a ogni riga di dati. Se si deseleziona questa casella di controllo vengono passati solo il campo ID record e i campi di modellazione in più, consentendo una più rapida esecuzione dello stream.

Calcola probabilità di classi assegnate per il punteggio dei record. (solo Albero decisionale e Bayes naive) Se selezionata, questa opzione indica che i campi di modellazione in più contengono un campo della confidenza (ovvero una probabilità) oltre al campo della previsione. Se si deseleziona questa casella di controllo viene generato solo il campo della previsione.

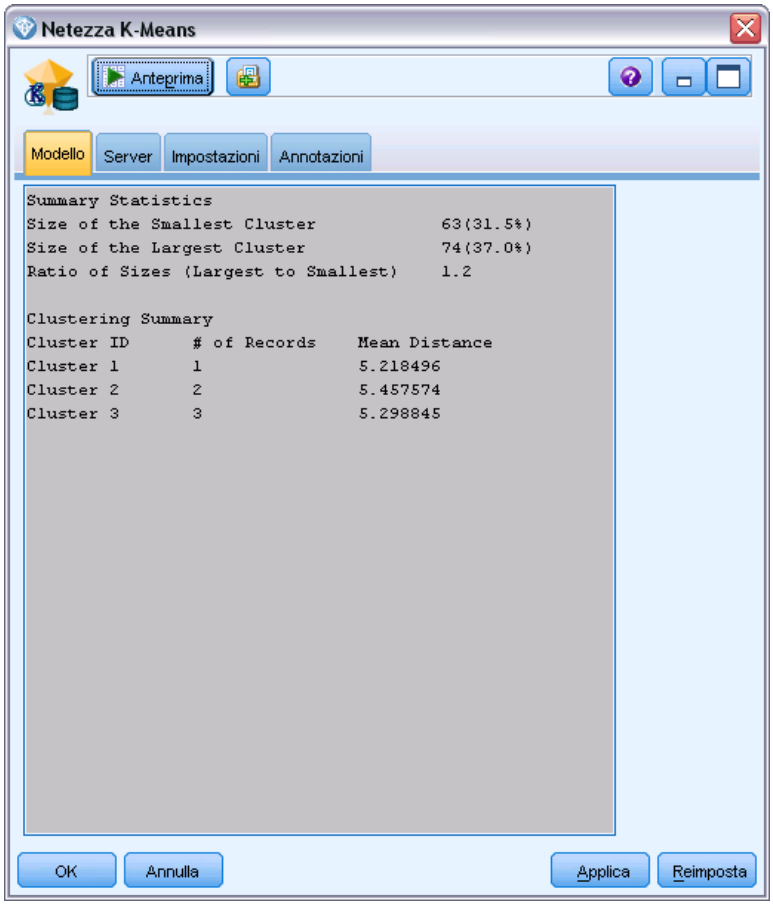
Insieme di modelli K-Means di Netezza

Gli insiemi di modelli K-Means contengono tutte le informazioni intercettate dal modello di cluster, nonché le informazioni sui dati di addestramento e sull'elaborazione della stima.

Quando si esegue uno stream che contiene un nodo Modelli K-Means, il nodo aggiunge due nuovi campi contenenti la classe di appartenenza e la distanza dal centro del 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 del cluster. Per esempio, se il modello è denominato *Kmeans*, i nuovi campi si chiameranno *$KM-Kmeans* e *$KMD-Kmeans*.

Insieme di modelli K-Means di Netezza - Scheda Modello

Figura 6-34
Output di un modello K-Means



L'output del modello è visualizzato nel modo seguente nella scheda Modello.

Statistiche riassuntive. Mostra il numero di record e la percentuale dell'insieme di dati occupata dai cluster sia per i cluster più grandi che per quelli più piccoli. L'elenco mostra inoltre il rapporto fra le dimensioni del cluster più grande e quelle del più piccolo.

Riepilogo cluster. Elenca i cluster creati dall'algoritmo. Per ogni cluster, la tabella mostra il numero di record e la loro distanza media dal centro del cluster.

Insieme di modelli K-Means di Netezza - Scheda Impostazioni

La scheda Impostazioni consente di impostare alcune opzioni di calcolo del punteggio del modello.

Includi campi di input. Se selezionata, questa opzione passa a valle tutti i campi di input originali, accodando il campo o i campi di modellazione in più a ogni riga di dati. Se si deseleziona questa casella di controllo vengono passati solo il campo ID record e i campi di modellazione in più, consentendo una più rapida esecuzione dello stream.

Misura della distanza. Metodo utilizzato per misurare la distanza tra i punti dei dati; una maggiore distanza indica una maggiore dissimilarità. Le opzioni disponibili sono:

- **Euclidea**. (default) La distanza fra due punti calcolata unendoli con una linea retta.
- **Manhattan**. La distanza fra due punti calcolata come somma delle differenze assolute fra le loro coordinate.
- **Canberra**. Simile alla distanza di Manhattan, ma più sensibile ai punti dei dati più vicini all'origine.
- **Massimo**. La distanza fra due punti calcolata come differenza massima sulla dimensione di qualsiasi coordinata.

Insiemi di modelli di rete di Bayes Netezza

L'insieme di modelli di rete di Bayes consente di impostare le opzioni per il calcolo del punteggio del modello.

Quando si esegue uno stream contenente un nodo Modelli di rete di Bayes, il nodo aggiunge un nuovo campo, il cui nome viene derivato dal nome del modello.

Tabella 6-7
Campo di calcolo del punteggio dei modelli per rete di Bayes

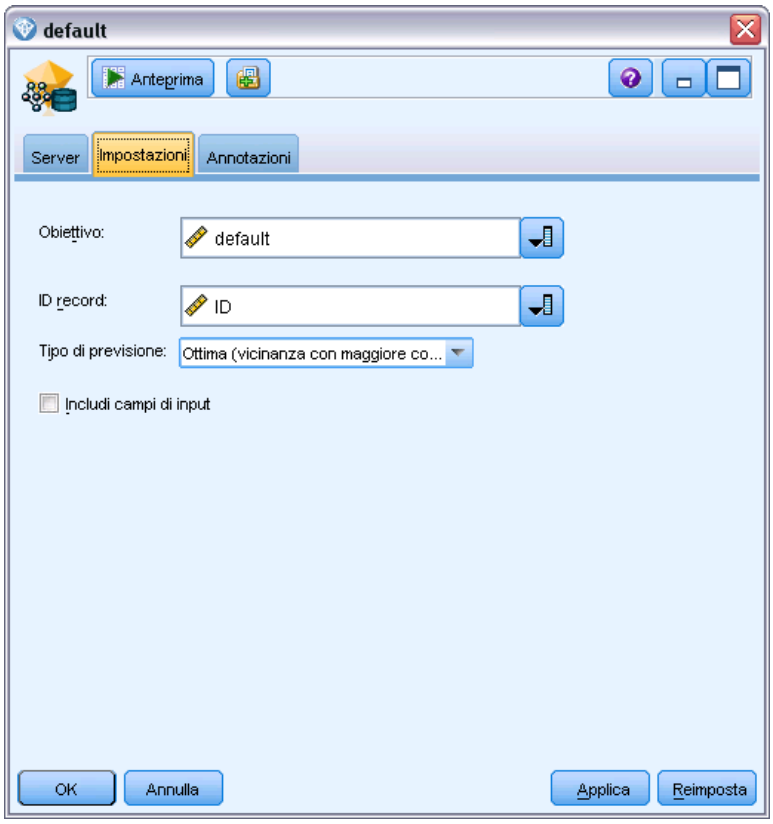
Nome del campo aggiunto	Significato
$BN-*nome_modello*	Valore previsto per il record corrente.

Per visualizzare il campo aggiuntivo, collegare un nodo Tabella all'insieme di modelli ed eseguire il nodo Tabella. Per ulteriori informazioni, vedere l'argomento Calcolo del punteggio dei modelli IBM Netezza Analytics a pag. 216.

Insieme di modelli di rete di Bayes Netezza - Scheda Impostazioni

Nella scheda Impostazioni è possibile impostare le opzioni per il calcolo del punteggio del modello.

Figura 6-35
Impostazioni del modello di rete di Bayes



Obiettivo. Se si desidera calcolare il punteggio di un campo obiettivo diverso dall'obiettivo corrente, scegliere qui il nuovo obiettivo.

ID record. Se non è specificato un campo ID record, scegliere qui il campo da utilizzare.

Tipo di previsione. La variazione dell'algoritmo di previsione da utilizzare:

- **Ottima (vicinanza con maggiore correlazione).** (default) Utilizza il nodo vicino con la maggiore correlazione.
- **Vicini (previsione ponderata dei vicini).** Utilizza una previsione ponderata di tutti i nodi vicini.
- **Vicini NN (vicini non nulli).** Equivale all'opzione precedente tranne per il fatto che ignora i nodi con valori nulli, ovvero i nodi corrispondenti agli attributi con valori mancanti per l'istanza per cui viene calcolata la previsione.

Includi campi di input. Se selezionata, questa opzione passa a valle tutti i campi di input originali, accodando il campo o i campi di modellazione in più a ogni riga di dati. Se si deseleziona questa casella di controllo vengono passati solo il campo ID record e i campi di modellazione in più, consentendo una più rapida esecuzione dello stream.

Insiemi di modelli Bayes naive Netezza

L'insieme di modelli Bayes naive consente di impostare le opzioni per il calcolo del punteggio del modello.

Quando si esegue uno stream contenente un nodo Modelli Bayes naive, il nodo aggiunge per default un nuovo campo, il cui nome viene derivato dal nome del modello.

Tabella 6-8
Campo di calcolo del punteggio dei modelli per Bayes naive - default

Nome del campo aggiunto	Significato
$I-*nome_modello*	Valore previsto per il record corrente.

Se si seleziona l'opzione Calcola probabilità di classi assegnate per il punteggio dei record nel nodo Modelli o nell'insieme di modelli e si esegue lo stream, vengono aggiunti altri due campi.

Tabella 6-9
Campi di calcolo del punteggio dei modelli per Bayes naive - aggiuntivi

Nome del campo aggiunto	Significato
$IP-*nome_modello*	Numeratore bayesiano della classe per l'istanza, ovvero il prodotto della probabilità della classe a priori e delle probabilità del valore dell'attributo dell'istanza condizionale.
$ILP-*nome_modello*	Algoritmo naturale del secondo elemento.

Per visualizzare i campi aggiuntivi, collegare un nodo Tabella all'insieme di modelli ed eseguire il nodo Tabella. Per ulteriori informazioni, vedere l'argomento Calcolo del punteggio dei modelli IBM Netezza Analytics a pag. 216.

Insieme di modelli Bayes naive Netezza - Scheda Impostazioni

Nella scheda Impostazioni è possibile impostare le opzioni per il calcolo del punteggio del modello.

Includi campi di input. Se selezionata, questa opzione passa a valle tutti i campi di input originali, accodando il campo o i campi di modellazione in più a ogni riga di dati. Se si deseleziona questa casella di controllo vengono passati solo il campo ID record e i campi di modellazione in più, consentendo una più rapida esecuzione dello stream.

Calcola probabilità di classi assegnate per il punteggio dei record. (solo Albero decisionale e Bayes naive) Se selezionata, questa opzione indica che i campi di modellazione in più contengono un campo della confidenza (ovvero una probabilità) oltre al campo della previsione. Se si deseleziona questa casella di controllo viene generato solo il campo della previsione.

- **Migliora la precisione della probabilità per insiemi di dati piccoli o notevolmente non bilanciati.** Quando si calcolano le probabilità, questa opzione richiama la tecnica di stima *m* per evitare le probabilità zero durante la stima. Questo tipo di stima delle probabilità può essere più lento ma fornisce risultati migliori per gli insiemi di dati di piccole dimensioni o notevolmente sbilanciati.

Insiemi di modelli KNN Netezza

L'insieme di modelli KNN consente di impostare le opzioni per il calcolo del punteggio del modello.

Quando si esegue uno stream contenente un nodo Modelli KNN, il nodo aggiunge un nuovo campo, il cui nome viene derivato dal nome del modello.

Tabella 6-10
Campo di calcolo del punteggio dei modelli per KNN

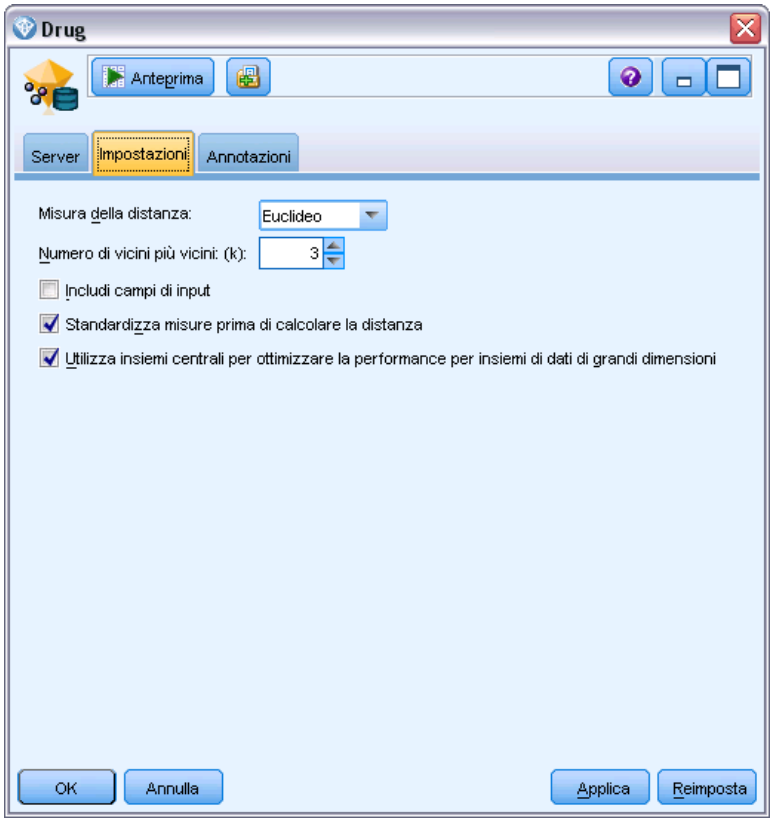
Nome del campo aggiunto	Significato
$KNN-*nome_modello*	Valore previsto per il record corrente.

Per visualizzare il campo aggiuntivo, collegare un nodo Tabella all'insieme di modelli ed eseguire il nodo Tabella. Per ulteriori informazioni, vedere l'argomento Calcolo del punteggio dei modelli IBM Netezza Analytics a pag. 216.

Insieme di modelli KNN Netezza - Scheda Impostazioni

Nella scheda Impostazioni è possibile impostare le opzioni per il calcolo del punteggio del modello.

Figura 6-36
Impostazioni del modello KNN



Misura della distanza. Metodo utilizzato per misurare la distanza tra i punti dei dati; una maggiore distanza indica una maggiore dissimilarità. Le opzioni disponibili sono:

- **Euclidea**. (default) La distanza fra due punti calcolata unendoli con una linea retta.
- **Manhattan**. La distanza fra due punti calcolata come somma delle differenze assolute fra le loro coordinate.
- **Canberra**. Simile alla distanza di Manhattan, ma più sensibile ai punti dei dati più vicini all'origine.
- **Massimo**. La distanza fra due punti calcolata come differenza massima sulla dimensione di qualsiasi coordinata.

Numero di Vicini più vicini (k). Il numero di vicini più vicini relativamente a un caso specifico. L'utilizzo di un numero maggiore di vicini non garantisce necessariamente un modello più preciso.

La scelta di *k* controlla la proporzione tra la prevenzione del sovradattamento (può essere importante, soprattutto per i dati "rumorosi") e la risoluzione (con previsioni diverse per istanze simili). Normalmente è necessario adattare il valore di *k* per ogni insieme di dati; i valori tipici variano da 1 a diverse decine.

Includi campi di input. Se selezionata, questa opzione passa a valle tutti i campi di input originali, accodando il campo o i campi di modellazione in più a ogni riga di dati. Se si deseleziona questa casella di controllo vengono passati solo il campo ID record e i campi di modellazione in più, consentendo una più rapida esecuzione dello stream.

Standardizza misure prima di calcolare la distanza. Se selezionata, questa opzione standardizza le misure per i campi di input continui prima di calcolare i valori della distanza.

Utilizza insiemi centrali per ottimizzare le performance per insiemi di dati di grandi dimensioni. Se selezionata, questa opzione utilizza il campionamento degli insiemi centrali per accelerare il calcolo quando si lavora con insiemi di dati di grandi dimensioni.

Insiemi di modelli di raggruppamento cluster divisivo Netezza

L'insieme di modelli di raggruppamento cluster divisivo consente di impostare le opzioni per il calcolo del punteggio del modello.

Quando si esegue uno stream contenente un nodo Modelli di raggruppamento cluster divisivo, il nodo aggiunge due nuovi campi, il cui nome viene derivato dal nome del modello.

Tabella 6-11
Campi di calcolo del punteggio dei modelli per raggruppamento cluster divisivo

Nome del campo aggiunto	Significato
$DC-*nome_modello*	Identificatore del sottocluster a cui viene assegnato il record corrente.
$DCD-*nome_modello*	Distanza dal centro del sottocluster per il record corrente.

Per visualizzare i campi aggiuntivi, collegare un nodo Tabella all'insieme di modelli ed eseguire il nodo Tabella. Per ulteriori informazioni, vedere l'argomento Calcolo del punteggio dei modelli IBM Netezza Analytics a pag. 216.

Insieme di modelli di raggruppamento cluster divisivo Netezza - Scheda Impostazioni

Nella scheda Impostazioni è possibile impostare le opzioni per il calcolo del punteggio del modello.

Includi campi di input. Se selezionata, questa opzione passa a valle tutti i campi di input originali, accodando il campo o i campi di modellazione in più a ogni riga di dati. Se si deseleziona questa casella di controllo vengono passati solo il campo ID record e i campi di modellazione in più, consentendo una più rapida esecuzione dello stream.

Misura della distanza. Metodo utilizzato per misurare la distanza tra i punti dei dati; una maggiore distanza indica una maggiore dissimilarità. Le opzioni disponibili sono:

- **Euclidea**. (default) La distanza fra due punti calcolata unendoli con una linea retta.
- **Manhattan**. La distanza fra due punti calcolata come somma delle differenze assolute fra le loro coordinate.
- **Canberra**. Simile alla distanza di Manhattan, ma più sensibile ai punti dei dati più vicini all'origine.
- **Massimo**. La distanza fra due punti calcolata come differenza massima sulla dimensione di qualsiasi coordinata.

Livello di gerarchia applicato. Livello della gerarchia da applicare ai dati.

Insiemi di modelli PCA Netezza

L'insieme di modelli PCA consente di impostare le opzioni per il calcolo del punteggio del modello.

Quando si esegue uno stream contenente un nodo Modelli PCA, il nodo aggiunge per default un nuovo campo, il cui nome viene derivato dal nome del modello.

Tabella 6-12
Campo di calcolo del punteggio dei modelli per PCA

Nome del campo aggiunto	Significato
$F-*nome_modello*	Valore previsto per il record corrente.

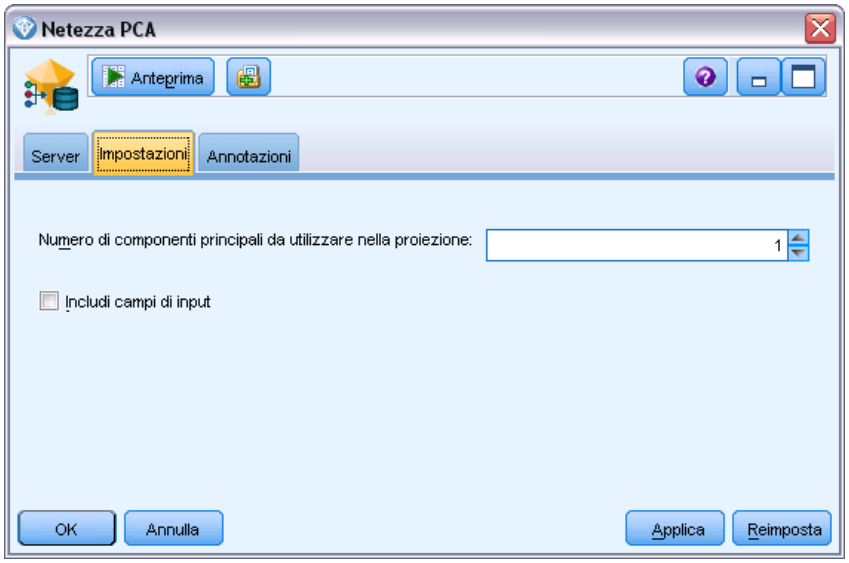
Se si specifica un valore maggiore di 1 nel campo Numero di componenti principali ... nel nodo Modelli o nell'insieme di modelli e si esegue lo stream, il nodo aggiunge un nuovo campo per ciascuna componente. In questo caso i nomi dei campi hanno il suffisso *-n*, dove *n* è il numero della componente. Per esempio, se il modello è denominato *pca* e contiene tre componenti, i nuovi campi saranno denominati *$F-pca-1*, *$F-pca-2* e *$F-pca-3*.

Per visualizzare i campi aggiuntivi, collegare un nodo Tabella all'insieme di modelli ed eseguire il nodo Tabella. Per ulteriori informazioni, vedere l'argomento Calcolo del punteggio dei modelli IBM Netezza Analytics a pag. 216.

Insieme di modelli PCA Netezza - Scheda Impostazioni

Nella scheda Impostazioni è possibile impostare le opzioni per il calcolo del punteggio del modello.

Figura 6-37
Impostazioni del modello PCA



Numero di componenti principali da utilizzare nella proiezione. Numero delle componenti principali a cui si desidera ridurre l'insieme di dati. Questo valore non deve superare il numero di attributi (campi di input).

Includi campi di input. Se selezionata, questa opzione passa a valle tutti i campi di input originali, accodando il campo o i campi di modellazione in più a ogni riga di dati. Se si deseleziona questa casella di controllo vengono passati solo il campo ID record e i campi di modellazione in più, consentendo una più rapida esecuzione dello stream.

Insiemi di modelli di albero di regressione Netezza

L'insieme di modelli di albero di regressione consente di impostare le opzioni per il calcolo del punteggio del modello.

Quando si esegue uno stream contenente un nodo Modelli dell'albero di regressione, il nodo aggiunge per default un nuovo campo, il cui nome viene derivato dal nome del modello.

Tabella 6-13
Campo di calcolo del punteggio dei modelli per l'albero di regressione

Nome del campo aggiunto	Significato
$I-*nome_modello*	Valore previsto per il record corrente.

Se si seleziona l'opzione Calcola varianza stimata nel nodo Modelli o nell'insieme di modelli e si esegue lo stream, viene aggiunto un ulteriore campo.

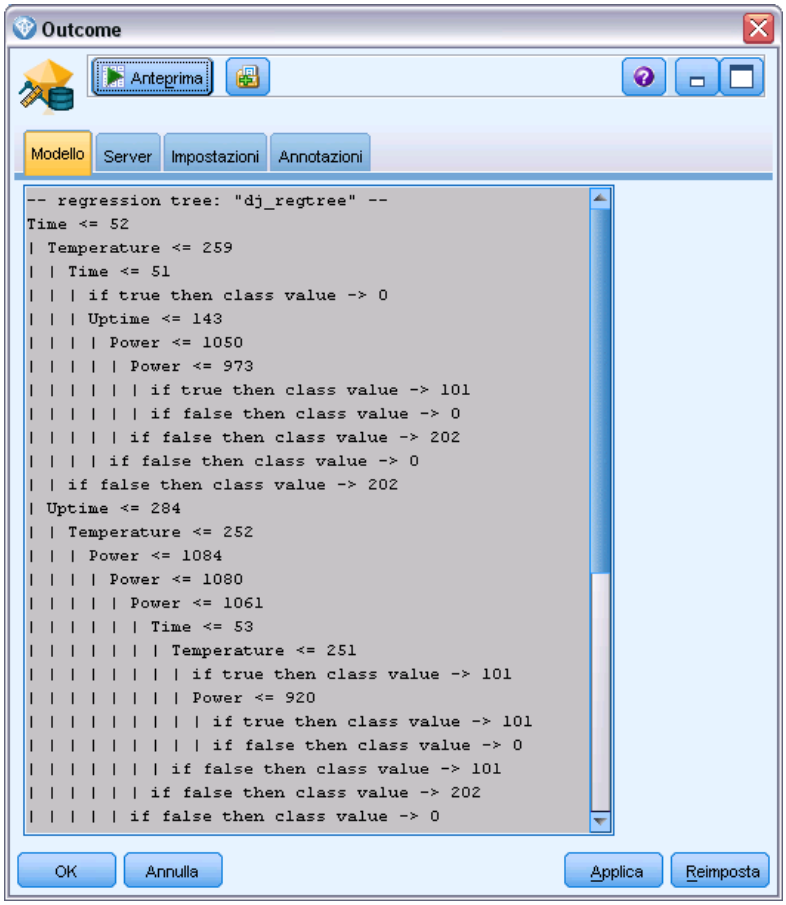
Tabella 6-14
Campo di calcolo del punteggio dei modelli per l'albero di regressione - aggiuntivo

Nome del campo aggiunto	Significato
$IV-*nome_modello*	Varianze stimate delle classi assegnate.

Per visualizzare i campi aggiuntivi, collegare un nodo Tabella all'insieme di modelli ed eseguire il nodo Tabella. Per ulteriori informazioni, vedere l'argomento Calcolo del punteggio dei modelli IBM Netezza Analytics a pag. 216.

Insieme di modelli dell'albero di regressione Netezza - Scheda Modello

Figura 6-38
Output modello di albero di regressione



L'output dei modelli assume la forma di una rappresentazione di testo dell'albero. Ogni riga di testo corrisponde a un nodo o una foglia e il rientro riflette il livello dell'albero. Per un nodo, viene visualizzata la condizione di suddivisione, per una foglia appare l'etichetta di classe assegnata.

Insieme di modelli dell'albero di regressione Netezza - Scheda Impostazioni

Nella scheda Impostazioni è possibile impostare le opzioni per il calcolo del punteggio del modello.

Includi campi di input. Se selezionata, questa opzione passa a valle tutti i campi di input originali, accodando il campo o i campi di modellazione in più a ogni riga di dati. Se si deseleziona questa casella di controllo vengono passati solo il campo ID record e i campi di modellazione in più, consentendo una più rapida esecuzione dello stream.

Calcola varianza stimata. Indica se le varianze delle classi assegnate devono essere incluse nell'output.

Insiemi di modelli di regressione lineare Netezza

L'insieme di modelli di regressione lineare consente di impostare le opzioni per il calcolo del punteggio del modello.

Quando si esegue uno stream contenente un nodo Modelli di regressione lineare, il nodo aggiunge un nuovo campo, il cui nome viene derivato dal nome del modello.

Tabella 6-15
Campo di calcolo del punteggio dei modelli per la regressione lineare

Nome del campo aggiunto	Significato
$LR-*nome_modello*	Valore previsto per il record corrente.

Insieme di modelli di regressione lineare Netezza - Scheda Impostazioni

Nella scheda Impostazioni è possibile impostare le opzioni per il calcolo del punteggio del modello.

Includi campi di input. Se selezionata, questa opzione passa a valle tutti i campi di input originali, accodando il campo o i campi di modellazione in più a ogni riga di dati. Se si deseleziona questa casella di controllo vengono passati solo il campo ID record e i campi di modellazione in più, consentendo una più rapida esecuzione dello stream.

Insieme di modelli di serie storica Netezza

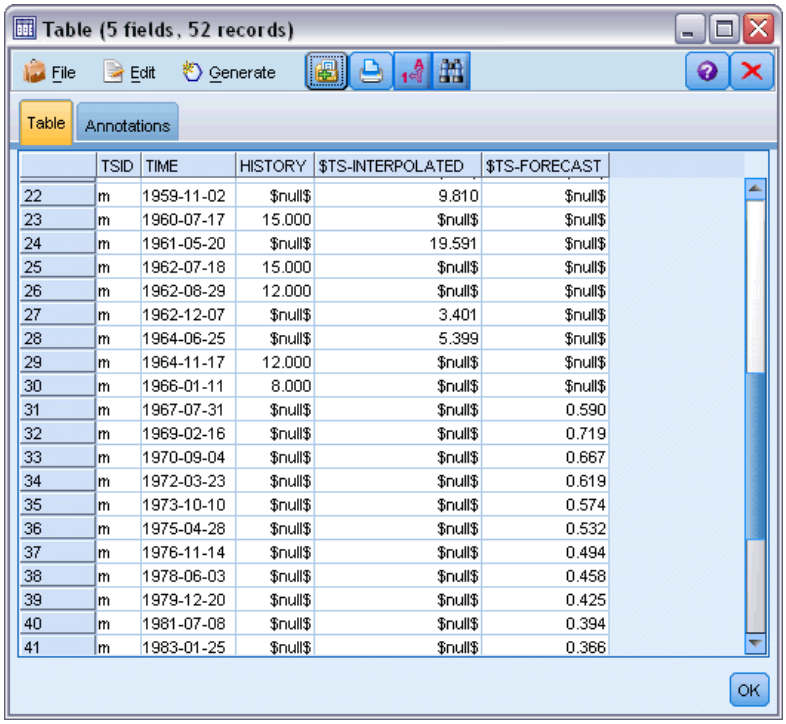
L'insieme di modelli consente di accedere all'output dell'operazione di creazione di modelli di serie storica. L'output è composto dai seguenti campi:

Tabella 6-16
Campi di output del modello di serie storica

Campo	Descrizione
TSID	L'identificatore della serie storica; il contenuto del campo ID serie storiche nella scheda Campi del nodo Modelli. Per ulteriori informazioni, vedere l'argomento Opzioni dei campi della serie storica Netezza a pag. 202.
TIME	Il periodo di tempo coperto dalla serie storica corrente.
HISTORY	I valori dei dati storici (utilizzati per effettuare la previsione). Il campo viene incluso solo se l'opzione Includi valori storici nel risultato è selezionata nella scheda Impostazioni dell'insieme di modelli.
$TS-INTERPOLATED	I valori interpolati, se presenti. Il campo viene incluso solo se l'opzione Includi valori interpolati nel risultato è selezionata nella scheda Impostazioni dell'insieme di modelli. Interpolazione è un'opzione della scheda Opzioni di creazione del nodo Modelli.
$TS-FORECAST	I valori di previsione per la serie storica.

Per visualizzare l'output del modello, allegare un nodo Tabella (dalla scheda Output della palette dei nodi) all'insieme di modelli ed eseguire il nodo Tabella. L'output tipico ha l'aspetto seguente.

Figura 6-39
Output tipico del modello di serie storica



	TSID	TIME	HISTORY	$TS-INTERPOLATED	$TS-FORECAST
22	m	1959-11-02	$null$	9.810	$null$
23	m	1960-07-17	15.000	$null$	$null$
24	m	1961-05-20	$null$	19.591	$null$
25	m	1962-07-18	15.000	$null$	$null$
26	m	1962-08-29	12.000	$null$	$null$
27	m	1962-12-07	$null$	3.401	$null$
28	m	1964-06-25	$null$	5.399	$null$
29	m	1964-11-17	12.000	$null$	$null$
30	m	1966-01-11	8.000	$null$	$null$
31	m	1967-07-31	$null$	$null$	0.590
32	m	1969-02-16	$null$	$null$	0.719
33	m	1970-09-04	$null$	$null$	0.667
34	m	1972-03-23	$null$	$null$	0.619
35	m	1973-10-10	$null$	$null$	0.574
36	m	1975-04-28	$null$	$null$	0.532
37	m	1976-11-14	$null$	$null$	0.494
38	m	1978-06-03	$null$	$null$	0.458
39	m	1979-12-20	$null$	$null$	0.425
40	m	1981-07-08	$null$	$null$	0.394
41	m	1983-01-25	$null$	$null$	0.366

Insieme serie storica Netezza - Scheda Impostazioni

Nella scheda Impostazioni è possibile specificare opzioni per personalizzare l'output del modello.

Nome modello. Il nome del modello specificato nella scheda Opzioni modello del nodo Modelli.

Le altre opzioni sono le stesse della scheda Opzioni di modellazione del nodo Modelli.

Insieme di modelli lineari generalizzati Netezza

L'insieme di modelli consente di accedere all'output dell'operazione di creazione di modelli.

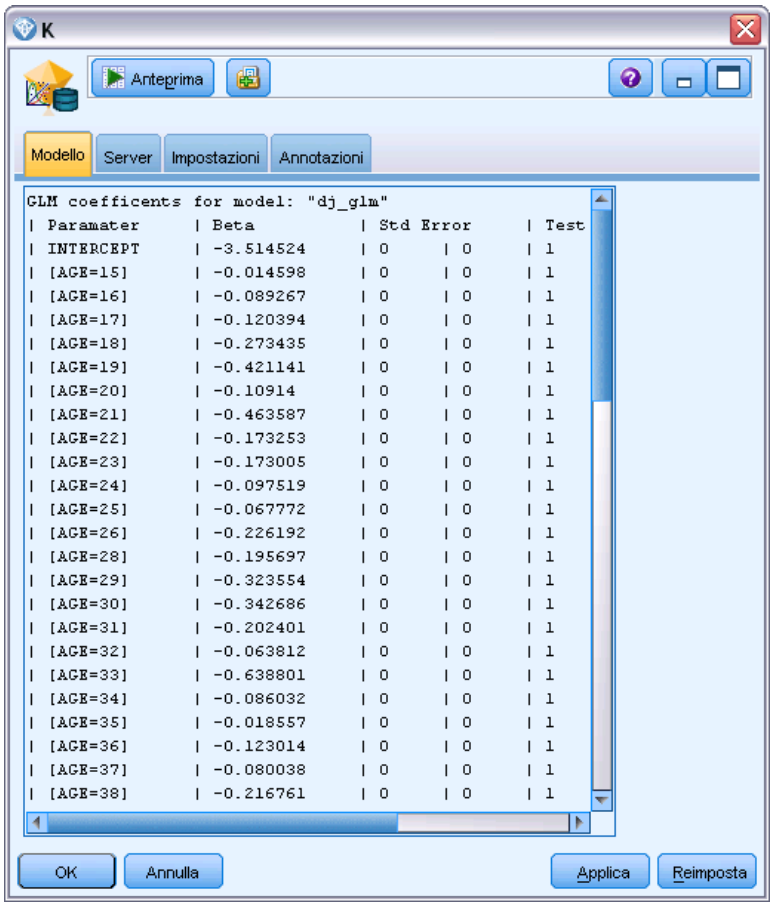
Quando si esegue uno stream contenente un nodo Modelli lineari generalizzati, il nodo aggiunge un nuovo campo, il cui nome viene derivato dal nome del modello.

Tabella 6-17
Campo di calcolo del punteggio dei modelli per lineare generalizzato

Nome del campo aggiunto	Significato
$GLM-*nome_modello*	Valore previsto per il record corrente.

La scheda Modello visualizza varie statistiche relative al modello.

Figura 6-40
Output del modello lineare generalizzato



L'output è composto dai seguenti campi:

Tabella 6-18
Campi di output del modello lineare generalizzato

Campo di output	Descrizione
Parametro	Parametri (vale a dire, le variabili predittore) utilizzati dal modello. Si tratta delle colonne numeriche e nominali, oltre all'intercetta (il termine costante nel modello di regressione).
Beta	Coefficiente di correlazione (vale a dire, il componente lineare del modello).
Errore std	Deviazione standard per beta.
Test	Statistiche di prova utilizzate per valutare la validità del parametro.
valore p	Probabilità di un errore quando si presuppone che il parametro sia significativo.
Riepilogo dei residui	
Tipo di residuo	Tipo di residuo della previsione per cui sono visualizzati i valori di riepilogo.
RSS	Valore del residuo.
df	Gradi di libertà per il residuo.
valore p	Probabilità di un errore. Un valore elevato indica un modello poco adatto; un valore basso indica un modello adatto.

Insieme di modelli lineari generalizzati Netezza - Scheda Impostazioni

Nella scheda Impostazioni è possibile personalizzare l'output del modello.

L'opzione corrisponde a quella descritta per Opzioni di calcolo del punteggio del nodo Modelli. Per ulteriori informazioni, vedere l'argomento Opzioni del modello lineare generalizzato Netezza - Opzioni di calcolo del punteggio a pag. 216.

Appendice

A

Note

Queste informazioni sono state preparate per prodotti e servizi offerti in tutto il mondo.

IBM potrebbe non offrire i prodotti, i servizi o le funzionalità di cui si tratta nel presente documento in altri paesi. Contattare il rappresentante IBM locale per informazioni sui prodotti e i servizi attualmente disponibili nella propria zona. Qualsiasi riferimento a un prodotto, programma o servizio IBM non intende dichiarare o implicare che sia possibile utilizzare esclusivamente tale prodotto, programma o servizio IBM. Potrà invece essere utilizzato qualsiasi prodotto, programma o servizio con funzionalità equivalente e che non violi i diritti di proprietà intellettuale di IBM. Tuttavia, è responsabilità dell'utente valutare e verificare il funzionamento di qualsiasi prodotto, programma o servizio non IBM.

IBM può essere titolare di brevetti o domande di brevetto relativi alla materia oggetto del presente documento. 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 Corporation, North Castle Drive, Armonk, NY 10504-1785, U.S.A.

Per richieste di informazioni sulle licenze riguardanti il set di caratteri a byte doppio (DBCS), contattare l'Intellectual Property Department di IBM del proprio paese, oppure inviare le richieste in forma scritta all'indirizzo:

Intellectual Property Licensing, Legal and Intellectual Property Law, IBM Japan Ltd., 1623-14, Shimotsuruma, Yamato-shi, Kanagawa 242-8502 Giappone.

Il seguente paragrafo non si applica per il Regno Unito o altri paesi in cui le presenti disposizioni non sono conformi alle leggi locali: INTERNATIONAL BUSINESS MACHINES FORNISCE QUESTA PUBBLICAZIONE "COSÌ COM'È" SENZA GARANZIA DI ALCUN TIPO, SIA ESSA ESPRESSA O IMPLICITA, INCLUSE, MA NON LIMITATE A, LE GARANZIE IMPLICITE DI NON VIOLAZIONE, COMMERCIABILITÀ O IDONEITÀ A UNO SCOPO SPECIFICO. 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 può apportare miglioramenti e/o modifiche al/ai prodotto/i e/o al/ai programma/i descritti nella presente pubblicazione in qualsiasi momento 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 contenuti in tali siti Web non fanno parte dei materiali di questo prodotto IBM e il loro utilizzo è esclusivamente a rischio dell'utente.

IBM può utilizzare o distribuire eventuali informazioni fornite dall'utente nei modi che ritiene appropriati senza incorrere in alcun obbligo nei confronti dell'utente.

I licenziatari del programma che desiderassero informazioni su di esso allo scopo di abilitare: (i) lo scambio di informazioni tra programmi creati indipendentemente e altri programmi (questo compreso) e (ii) l'utilizzo in comune delle informazioni scambiate, dovranno rivolgersi a:

IBM Software Group, All'attenzione di: Licensing, 233 S. Wacker Dr., Chicago, IL 60606, USA.

Tali informazioni saranno fornite in conformità ai termini e alle condizioni in vigore e, in alcuni casi, dietro pagamento.

Il programma concesso in licenza descritto nel presente documento e tutto il materiale correlato disponibile sono forniti da IBM in base ai termini del contratto di licenza cliente IBM, del contratto di licenza internazionale IBM o del contratto equivalente esistente 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 verificato tali prodotti e non può confermare l'accuratezza delle prestazioni, la compatibilità o qualsiasi altra dichiarazione relativa a prodotti non IBM. Eventuali domande in merito alle funzionalità dei prodotti non IBM vanno indirizzate ai fornitori di tali prodotti.

Qualsiasi 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 aziendali 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, ibm.com e SPSS sono marchi di IBM Corporation, registrati in numerose giurisdizioni nel mondo. Un elenco aggiornato dei marchi IBM è disponibile sul Web all'indirizzo *http://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 i logo basati su Java sono marchi di Sun Microsystems, Inc. negli Stati Uniti e/o negli altri paesi.

Altri nomi di prodotti e servizi possono essere marchi commerciali di IBM o di altre aziende.

Indice