IBM SPSS Modeler 15 Guida alle applicazioni

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*.

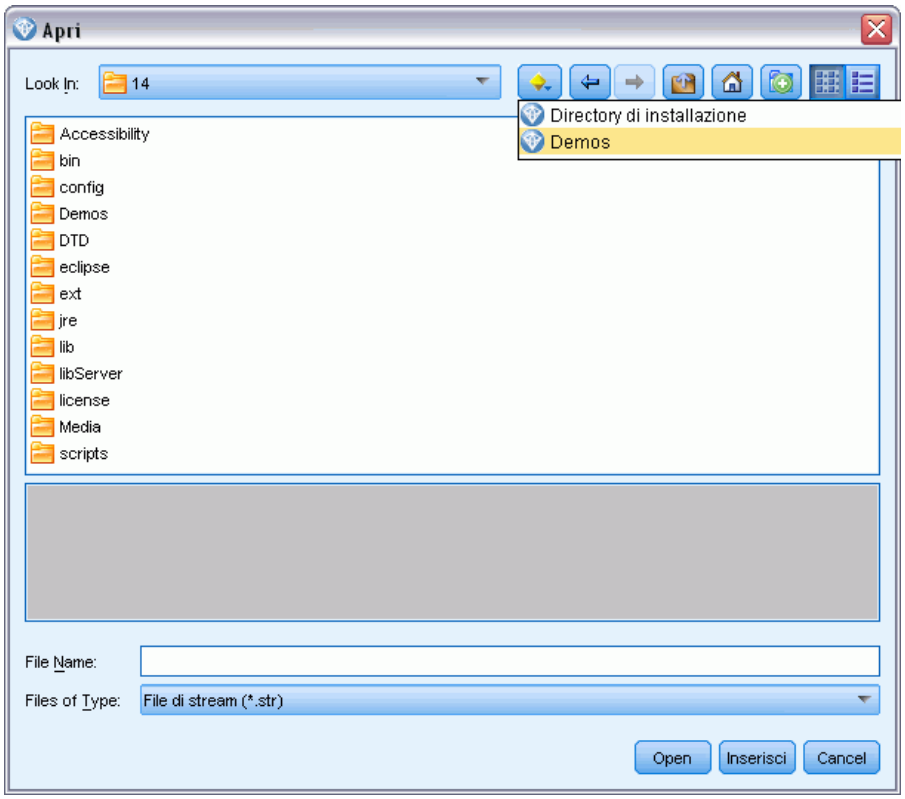
Esempi di modellazione in-database. Vedere gli esempi nella *IBM SPSS Modeler Guida alla modellazione in-database*.

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



Parte I:
Introduzione e Guida introduttiva

Capitolo
2

Panoramica su IBM SPSS Modeler

Guida Introduttiva

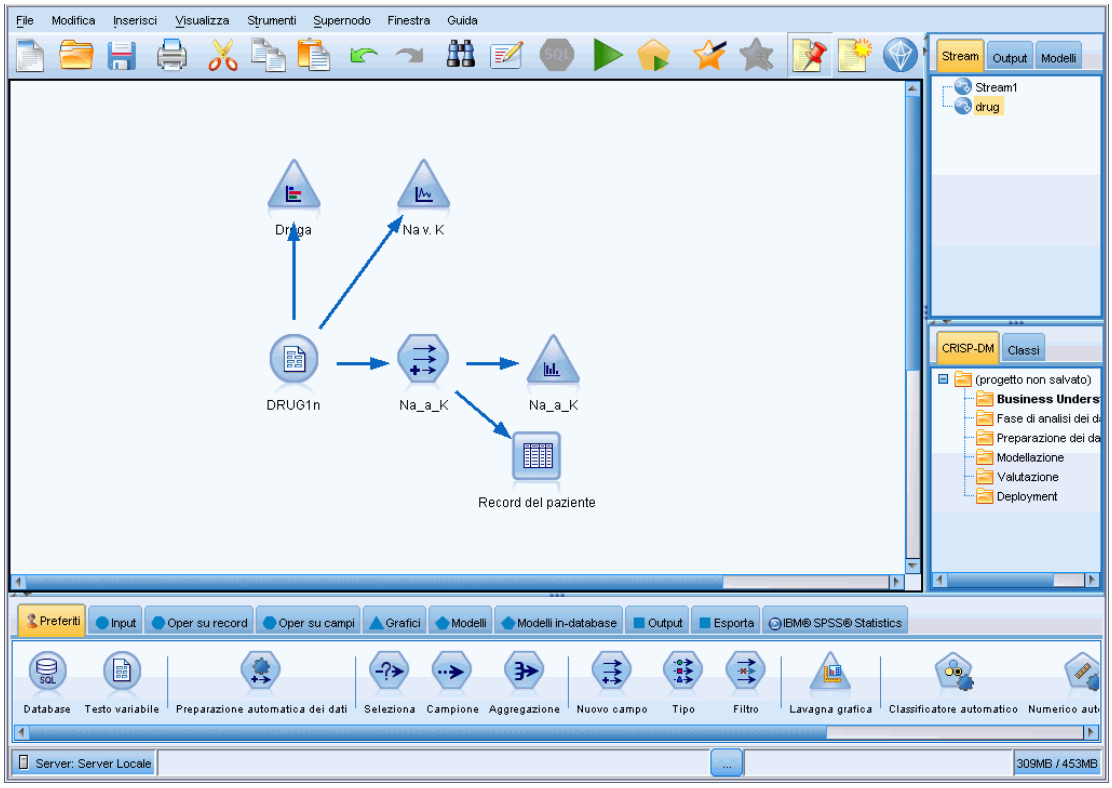
IBM® SPSS® Modeler è un'applicazione di data mining che offre un approccio strategico al rilevamento di relazioni utili all'interno di insiemi di dati di grandi dimensioni. Rispetto ai metodi statistici più tradizionali, non è necessario conoscere fin dall'inizio che cosa cercare esattamente. È possibile esplorare i dati, adattarvi modelli diversi e analizzare le diverse relazioni, fino a individuare informazioni utili.

Avvio di IBM SPSS Modeler

Per avviare l'applicazione, fare clic su
Start > [Tutti] i programmi > IBM SPSS Modeler15 > IBM SPSS Modeler15

Dopo qualche secondo viene visualizzata la finestra principale.

Figura 2-1
Finestra principale di IBM SPSS Modeler



Avvio dalla riga di comando

È possibile utilizzare la riga di comando del sistema operativo per avviare IBM® SPSS® Modeler:

- Sul computer in cui è installato IBM® SPSS® Modeler, aprire una finestra DOS (prompt dei comandi).
- Per avviare l'interfaccia di SPSS Modeler in modalità interattiva, digitare il comando modelerclient seguito dagli argomenti desiderati; per esempio:

```
modelerclient -stream report.str -execute
```

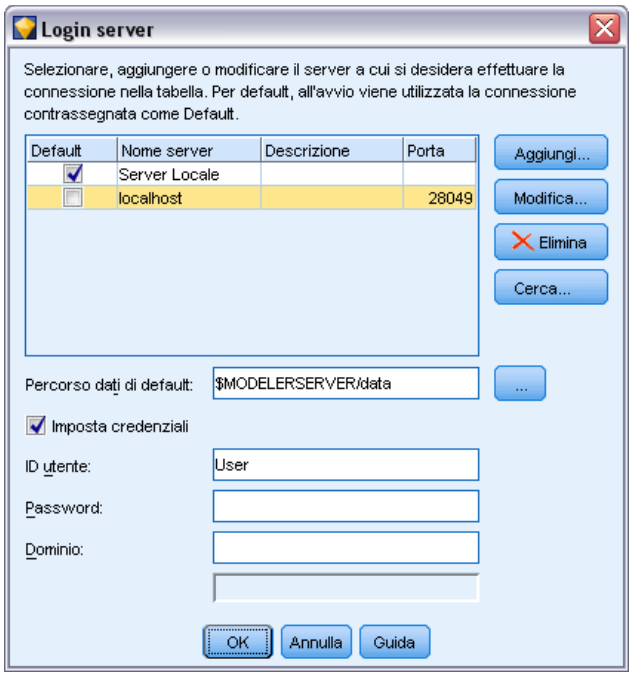
Gli argomenti disponibili (flag) consentono di connettersi a un server, caricare stream, eseguire script o specificare altri parametri.

Connessione a IBM SPSS Modeler Server

È possibile eseguire IBM® SPSS® Modeler come applicazione autonoma oppure come client connesso direttamente a IBM® SPSS® Modeler Server o a SPSS Modeler Server o a un cluster di server tramite il plug-in Coordinator of Processes disponibile in IBM® SPSS® Collaboration and Deployment Services. Lo stato di connessione corrente è visualizzato nella parte inferiore sinistra della finestra di SPSS Modeler.

Per connettersi a un server, è possibile immettere manualmente il nome del server al quale connettersi oppure selezionare un nome definito in precedenza. Tuttavia, se si dispone di IBM SPSS Collaboration and Deployment Services, è possibile cercare nell'elenco di server o cluster di server disponibili nella finestra di dialogo Login server. Il plug-in Coordinator of Processes consente di spostarsi tra i servizi Statistics in esecuzione in una rete. Per ulteriori informazioni, vedere l'argomento Bilanciamento del carico con cluster di server in l'appendice D in *IBM SPSS Modeler Server 15 Guida della performance e amministrazione*.

Figura 2-2
Finestra di dialogo Login server



Per connettersi a un server

- Nel menu Strumenti, fare clic su Login server. Viene visualizzata la finestra di dialogo Login server. In alternativa, fare doppio clic sull'area dello stato di connessione nella finestra di SPSS Modeler.
- Nella finestra di dialogo specificare le opzioni per la connessione al server locale oppure selezionare una connessione dalla tabella.
  - Fare clic su Aggiungi o Modifica per aggiungere o modificare una connessione. Per ulteriori informazioni, vedere l'argomento Aggiunta e modifica della connessione di IBM SPSS Modeler Server in *Manuale dell'utente di IBM SPSS Modeler 15*.
  - Fare clic su Cerca per accedere al server o al cluster di server in Coordinator of Processes. Per ulteriori informazioni, vedere l'argomento Ricerca di server in IBM SPSS Collaboration and Deployment Services in *Manuale dell'utente di IBM SPSS Modeler 15*.

Tabella Server. Questa tabella contiene l'insieme di connessioni server definite. Nella tabella sono visualizzati la connessione di default, il nome del server, la descrizione e il numero di porta. È possibile aggiungere manualmente una nuova connessione, nonché selezionare o cercare una connessione esistente. Per impostare un determinato server come connessione di default, selezionare la casella di controllo nella colonna Default della tabella per la connessione desiderata.

Percorso dati di default. Specificare il percorso utilizzato per i dati nel server. Fare clic sul pulsante con i puntini di sospensione (...) per passare alla posizione richiesta.

Imposta credenziali. Lasciare questa casella deselezionata per abilitare la funzione di **Single Sign-On**, che cercherà di far accedere l'utente al server utilizzando il nome utente e la password del computer locale. Se la funzione di Single Sign-On non è disponibile oppure se si seleziona la casella per disabilitare Single Sign-On (per esempio, per accedere a un account amministratore), vengono attivati i seguenti campi per l'inserimento delle credenziali.

ID utente. Immettere il nome utente con il quale accedere al server.

Password. Immettere la password associata al nome utente specificato.

Dominio. Specificare il dominio utilizzato per accedere al server. È necessario specificare un nome di dominio solo quando il computer server risiede in un dominio Windows diverso da quello del computer client.

- Fare clic su OK per completare la connessione.

Per disconnettersi da un server

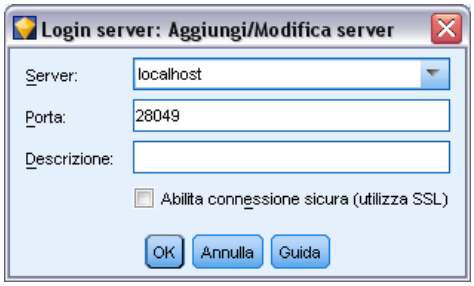
- Nel menu Strumenti, fare clic su Login server. Viene visualizzata la finestra di dialogo Login server. In alternativa, fare doppio clic sull'area dello stato di connessione nella finestra di SPSS Modeler.
- Nella finestra di dialogo, selezionare il Server locale e fare clic su OK.

Aggiunta e modifica della connessione di IBM SPSS Modeler Server

È possibile modificare o aggiungere manualmente una connessione server nella finestra di dialogo Login server. Facendo clic su Aggiungi, è possibile accedere a una finestra di dialogo Aggiungi/Modifica server vuota in cui è possibile immettere i dettagli di connessione al server. Dopo aver selezionato una connessione esistente e aver fatto clic su Modifica nella finestra di dialogo Login server, viene visualizzata la finestra di dialogo Aggiungi/Modifica server con i dettagli della connessione da modificare.

Nota: non è possibile modificare una connessione server che è stata aggiunta da IBM® SPSS® Collaboration and Deployment Services, poiché il nome, la porta e altri dettagli sono definiti in IBM SPSS Collaboration and Deployment Services.

Figura 2-3
Finestra di dialogo Login server: Aggiungi/Modifica server



Per aggiungere connessioni server

- Nel menu Strumenti, fare clic su Login server. Viene visualizzata la finestra di dialogo Login server.

- Nella finestra di dialogo, fare clic su Aggiungi. Viene visualizzata la finestra di dialogo Login server: Aggiungi/Modifica server.

- Immettere i dettagli della connessione server e fare clic su OK per salvare la connessione e ritornare alla finestra di dialogo Login server.
  - **Server.** Specificare uno dei server disponibili o selezionarne uno dall'elenco. È possibile identificare il computer server tramite un nome alfanumerico, per esempio *serverpersonale*, oppure tramite l'indirizzo IP assegnato al computer server, per esempio 202.123.456.78.
  - **Porta.** Indica il numero di porta su cui il server rimane in ascolto. Se l'impostazione predefinita non funziona, chiedere il numero di porta corretto all'amministratore di sistema.
  - **Descrizione.** Immettere una descrizione facoltativa della connessione al server.
  - **Abilita connessione sicura (utilizza SSL).** Specifica se è necessario utilizzare una connessione SSL (**Secure Sockets Layer**). SSL è un protocollo molto diffuso per la gestione della sicurezza dei dati trasmessi in rete. Per utilizzare questa funzione è necessario che il protocollo SSL sia stato attivato sul server di hosting di IBM® SPSS® Modeler Server. Se sono necessarie ulteriori informazioni, rivolgersi all'amministratore di sistema.

Per modificare connessioni server

- Nel menu Strumenti, fare clic su Login server. Viene visualizzata la finestra di dialogo Login server.

- Nella finestra di dialogo, selezionare la connessione da modificare, quindi fare clic su Modifica. Viene visualizzata la finestra di dialogo Login server: Aggiungi/Modifica server.

- Modificare i dettagli della connessione server e fare clic su OK per salvare le modifiche e ritornare alla finestra di dialogo Login server.

Ricerca di server in IBM SPSS Collaboration and Deployment Services

Anziché immettere una connessione server manualmente, è possibile selezionare un server o un cluster di server della rete tramite Coordinator of Processes, disponibile in IBM® SPSS® Collaboration and Deployment Services. Un cluster di server è un gruppo di server tra i quali Coordinator of Processes stabilisce quello più adatto a rispondere a una richiesta di elaborazione. Per ulteriori informazioni, vedere l'argomento Bilanciamento del carico con cluster di server in l'appendice D in *IBM SPSS Modeler Server 15 Guida della performance e amministrazione*.

Sebbene nella finestra di dialogo Login server sia possibile aggiungere server manualmente, cercando i server disponibili è possibile connettersi agli stessi senza doverne conoscere il nome e il numero di porta. Queste informazioni sono infatti fornite automaticamente. È in ogni caso necessario disporre delle informazioni di accesso corrette, quali nome utente, dominio e password.

Nota: se non si dispone della funzionalità Coordinator of Processes, è comunque possibile immettere manualmente il nome del server a cui connettersi oppure selezionare un nome definito in precedenza. Per ulteriori informazioni, vedere l'argomento Aggiunta e modifica della connessione di IBM SPSS Modeler Server in *Manuale dell'utente di IBM SPSS Modeler 15*.

Figura 2-4
Finestra di dialogo Cerca server

Per cercare server e cluster

- Nel menu Strumenti, fare clic su Login server. Viene visualizzata la finestra di dialogo Login server.
- Nella finestra di dialogo, fare clic su Cerca per aprire la finestra di dialogo Cerca server. Se non si è ancora connessi a IBM SPSS Collaboration and Deployment Services quando si tenta di sfogliare Coordinator of Processes, viene richiesto di connettersi. Per ulteriori informazioni, vedere l'argomento Connessione al repository in il capitolo 9 in *Manuale dell'utente di IBM SPSS Modeler 15*.
- Selezionare un server o un cluster di server dall'elenco.
- Fare clic su OK per chiudere la finestra di dialogo e aggiungere la connessione alla tabella nella finestra di dialogo Login server.

Modifica della directory Temp

Alcune operazioni eseguite da IBM® SPSS® Modeler Server potrebbero richiedere la creazione di file temporanei. Per default, per la creazione di file temporanei in IBM® SPSS® Modeler viene utilizzata la directory temporanea di sistema. È possibile modificare il percorso della directory temporanea, attenendosi alla procedura seguente.

- Creare una nuova directory con nome *spss* e una sottodirectory con nome *servertemp*.
- Modificare il file *options.cfg*, disponibile nella directory */config* dell'installazione di SPSS Modeler in uso. Modificare il parametro temp_directory in tale file come segue: temp_directory, "C:/spss/servertemp".
- Dopo avere eseguito tale operazione, è necessario riavviare il servizio SPSS Modeler Server, facendo clic sulla scheda Servizi nel Pannello di controllo di Windows. Arrestare il servizio, quindi avviarlo per attivare le modifiche apportate. È possibile riavviare il servizio anche riavviando il computer.

  Tutti i file temporanei verranno quindi scritti nella nuova directory.

Nota: l'errore più comune in questo tipo di operazione è l'utilizzo del tipo di barra non corretto. In conformità con UNIX, in SPSS Modeler vengono utilizzate le barre (/).

Avvio di più sessioni di IBM SPSS Modeler

Per avviare più sessioni di IBM® SPSS® Modeler contemporaneamente occorre apportare alcune modifiche alle impostazioni di IBM® SPSS® Modeler e di Windows. Questo può essere necessario, per esempio, se si dispone di due licenze server separate e si desidera eseguire due stream su due server diversi dallo stesso computer client.

Per attivare sessioni multiple di SPSS Modeler:

- Fare clic su
  Start > [Tutti] i programmi > IBM SPSS Modeler15
- Con il pulsante destro del mouse, fare clic sul collegamento IBM SPSS Modeler15 (quello con l'icona) e selezionare Proprietà.
- Nella casella di testo Destinazione, aggiungere -noshare alla fine della stringa.
- In Esplora risorse, selezionare:
  Strumenti > Opzioni cartella...
- Nella scheda Tipi di file, selezionare l'opzione Stream di SPSS Modeler e fare clic su Avanzate.
- Nella finestra di dialogo Modifica tipo file, selezionare Open with SPSS Modeler e fare clic su Modifica.
- Nella casella di testo Applicazione utilizzata per eseguire l'operazione, aggiungere -noshare prima dell'argomento -stream.

Nozioni di base sull'interfaccia di IBM SPSS Modeler

In ogni fase del processo di data mining, l'interfaccia visiva di IBM® SPSS® Modeler sollecita le capacità commerciali specifiche dell'utente. Gli algoritmi di creazione di modelli, quali quelli di previsione, classificazione, segmentazione e individuazione delle associazioni, assicurano modelli potenti e precisi. I modelli possono essere facilmente implementati e letti nei database, nell'IBM® SPSS® Statistics e in molte altre applicazioni.

Il processo di elaborazione dei dati in SPSS Modeler è articolato in tre passaggi.

- In primo luogo, si leggono i dati in SPSS Modeler.
- Quindi li si sottopone a una serie di manipolazioni.
- Infine, li si invia a una destinazione.

Questa sequenza di operazioni è nota come **stream di dati** perché i dati fluiscono, record dopo record, dalla sorgente attraverso ogni manipolazione e infine arrivano alla destinazione, che può essere un modello o un tipo di output di dati.

Figura 2-5
Uno stream semplice

Area di disegno dello stream di IBM SPSS Modeler

L'area di disegno dello stream è l'area di maggiori dimensioni della finestra di IBM® SPSS® Modeler. In quest'area vengono costruiti e manipolati gli stream di dati.

Gli stream vengono creati disegnando diagrammi delle operazioni sui dati pertinenti per la propria azienda nell'area di disegno principale dell'interfaccia. Ogni operazione è rappresentata da un'icona o **nodo** e i nodi sono collegati insieme in uno **stream** che rappresenta il flusso dei dati attraverso ciascuna operazione.

Con SPSS Modeler è possibile utilizzare più stream contemporaneamente, nella stessa area di disegno o aprendo una nuova area di disegno dello stream. Nel corso di una sessione, gli stream vengono archiviati nel relativo Manager, nella parte superiore destra della finestra di SPSS Modeler.

Palette nodi

La maggior parte dei dati e degli strumenti di modellazione di IBM® SPSS® Modeler si trovano nella **palette dei nodi**, visualizzata nella sezione sottostante l'area di disegno dello stream.

Per esempio, la scheda Oper su record della palette contiene i nodi utilizzabili per eseguire operazioni sui **record** di dati quali la selezione, l'unione e l'accodamento.

Per aggiungere nodi all'area di disegno, fare doppio clic sulle icone dalla palette dei nodi oppure trascinare e rilasciare le icone nell'area di disegno. Connettere quindi le icone per creare uno **stream** che rappresenti il flusso di dati.

Figura 2-6
Scheda Oper su record della palette dei nodi

Ogni scheda della palette contiene un insieme di nodi correlati che viene utilizzato per le diverse fasi delle operazioni degli stream, per esempio:

- **Input.** Nodi che consentono di inserire i dati in SPSS Modeler.
- **Oper su record.** Nodi utilizzati per le operazioni sui **record** di dati, per esempio la selezione, l'unione e l'accodamento.
- **Oper su campi.** Nodi utilizzati per le operazioni sui **campi** di dati, per esempio l'applicazione di filtri, la derivazione di nuovi campi e la definizione del livello di misurazione per campi specifici.
- **Grafici.** Nodi utilizzati per rappresentare graficamente i dati prima e dopo la fase di modellazione. I grafici includono i plot, gli istogrammi, i nodi Web e i grafici di valutazione.
- **Creazione di modelli.** Nodi che utilizzano gli algoritmi di creazione di modelli disponibili in SPSS Modeler, per esempio le reti neurali, gli alberi decisionali, gli algoritmi cluster e la sequenzializzazione dei dati.
- **Modelli in-database.** Nodi che utilizzano gli algoritmi di modellazione disponibili nei database di Microsoft SQL Server, IBM DB2 Oracle.
- **Output.** Nodi che generano output di svariati tipi per dati, grafici e risultati di modelli, visualizzabili in SPSS Modeler.
- **Esporta.** Nodi che generano vari tipi di output, visualizzabili in applicazioni esterne quali IBM® SPSS® Data Collection o Excel.
- **SPSS Statistics.** Nodi utilizzati per importare o esportare dati da IBM® SPSS® Statistics e per eseguire le procedure di SPSS Statistics.

Una volta acquisita una certa dimestichezza con le funzioni di SPSS Modeler, sarà possibile personalizzare il contenuto della palette a piacere. Per ulteriori informazioni, vedere l'argomento Personalizzazione della palette dei nodi in il capitolo 12 in *Manuale dell'utente di IBM SPSS Modeler 15*.

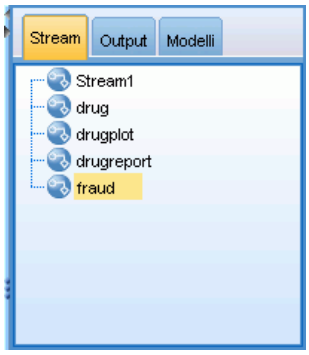
Il riquadro Report, posto sotto la palette dei nodi, consente di controllare l'andamento delle varie operazioni, per esempio in fase di lettura dei dati nello stream di dati. Anche il riquadro dello stato si trova sotto la palette dei nodi. Esso fornisce informazioni sulle operazioni in corso nell'applicazione e offre indicazioni qualora sia richiesto il feedback dell'utente.

Manager di IBM SPSS Modeler

Il riquadro dei manager si trova nella parte superiore destra della finestra ed è formato da tre schede, utilizzate per gestire gli stream, l'output e i modelli.

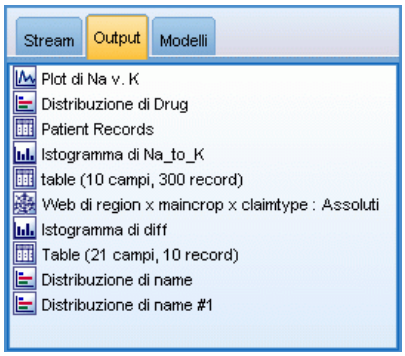
La scheda Stream può essere utilizzata per aprire, rinominare, salvare ed eliminare gli stream creati in una sessione.

Figura 2-7
Scheda Stream



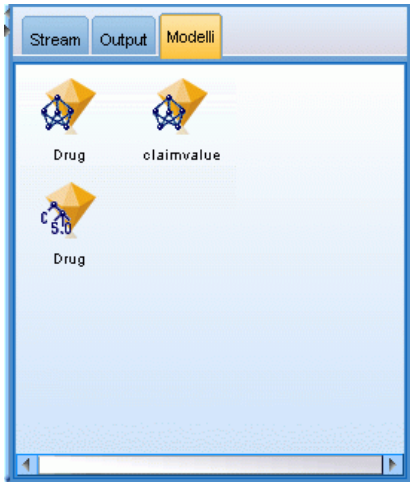
La scheda Output contiene tutti i file creati (quali grafici e tabelle) tramite le operazioni su stream in IBM® SPSS® Modeler. I grafici, le tabelle e i report elencati in questa scheda possono essere visualizzati, salvati, rinominati e chiusi.

Figura 2-8
Scheda Output



La scheda Modelli è la scheda più importante della finestra dei manager. Essa contiene tutti gli **insiemi di modelli**, che contengono i modelli generati in SPSS Modeler per la sessione corrente. I modelli possono essere visualizzati direttamente dalla scheda Modelli oppure possono essere aggiunti allo stream nell'area di disegno.

Figura 2-9
Scheda Modelli contenente degli insiemi di modelli

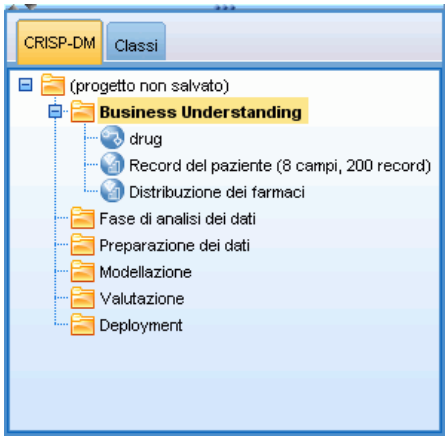


Progetti di IBM SPSS Modeler

Nella parte inferiore destra della finestra viene visualizzato il riquadro dei progetti, utilizzato per creare e gestire **progetti** di data mining (gruppi di file relativi a un'attività di data mining). Sono disponibili due modalità di visualizzazione per i progetti creati in IBM® SPSS® Modeler—: le visualizzazioni Classi e CRISP-DM.

La scheda CRISP-DM consente di organizzare i progetti in base allo standard CRISP-DM (Cross-Industry Standard Process for Data Mining), una metodologia del settore consolidata e non proprietaria. L'utilizzo dello strumento CRISP-DM consente ai data miner esperti e principianti di organizzare e comunicare al meglio le operazioni eseguite.

Figura 2-10
visualizzazione CRISP-DM



La scheda Classi consente di organizzare il proprio lavoro in SPSS Modeler per categorie, in base ai tipi di oggetti creati. Si tratta di una visualizzazione utile per l'inventario di dati, stream e modelli.

Figura 2-11
visualizzazione Classi

IBM SPSS Modeler Barra degli strumenti

Nella parte superiore della finestra di IBM® SPSS® Modeler vi è una barra degli strumenti a icone che offre diverse funzioni utili. Di seguito vengono illustrati i pulsanti della barra degli strumenti e le relative funzioni.


Crea un nuovo stream


Apri stream


Salva stream


Stampa stream corrente


Taglia e sposta negli Appunti


Copia negli Appunti


Incollare la selezione


Annulla l'ultima operazione


Ripeti


Cerca nodi


Modifica proprietà dello stream

Anteprima generazione SQL

	Esegui stream corrente		Esegui selezione di stream
	Interrompi stream (attivato solo durante l'esecuzione dello stream)		Aggiunge un Supernodo
	Zoom avanti (solo Supernodi)		Zoom indietro (solo Supernodi)
	No markup in stream		Inserisci commento
	Nascondi markup stream (se presente)		Mostra markup dello stream nascosto
	Apri stream in IBM® SPSS® Modeler Advantage		

Il markup dello stream comprende i commenti agli stream, i collegamenti dei modelli e l'indicazione dei rami di calcolo del punteggio.

Per maggiori informazioni sui commenti agli stream, vedere Aggiunta di commenti e annotazioni a nodi e stream a pag. .

Per maggiori informazioni sull'indicazione dei rami di calcolo del punteggio, vedere Il ramo di calcolo del punteggio a pag. .

I collegamenti dei modelli sono descritti nel manuale *Nodi Modelli in IBM SPSS*.

Personalizzazione della barra degli strumenti

È possibile modificare vari aspetti della barra degli strumenti, per esempio:

- Se è visualizzata o meno
- Se viene visualizzato il testo descrittivo delle icone
- Se utilizzare le icone grandi o piccole

Per attivare o disattivare la visualizzazione della barra degli strumenti:

- Nel menu principale, fare clic su:
  Visualizza > Barra degli strumenti > Visualizza

Per modificare l'impostazione del testo descrittivo o delle dimensioni delle icone:

- Nel menu principale, fare clic su:
  Visualizza > Barra degli strumenti > Personalizza

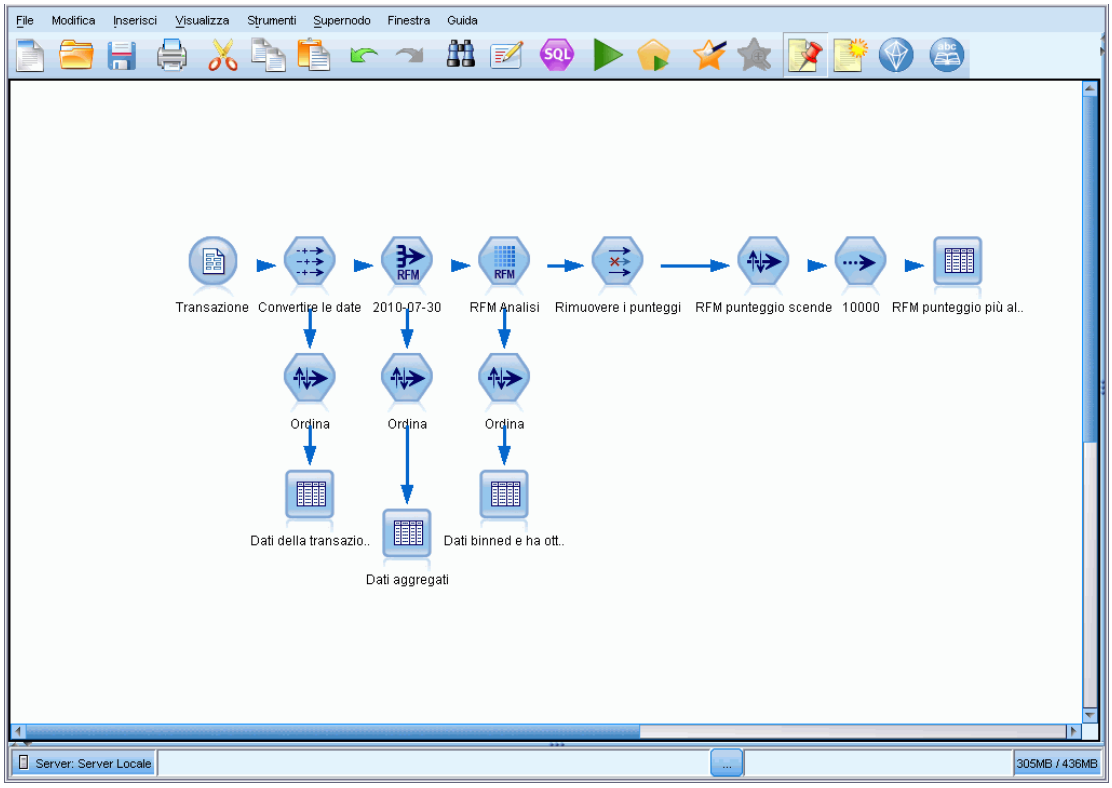
Fare clic su Mostra descrizioni o Pulsanti grandi a seconda dei casi.

Personalizzazione della finestra di IBM SPSS Modeler

I divisori tra le varie parti dell'interfaccia di IBM® SPSS® Modeler consentono di ridimensionare o chiudere a piacere gli strumenti. Per esempio, per uno stream di grandi dimensioni è possibile utilizzare le piccole frecce poste su ciascun divisore per chiudere la palette dei nodi, il riquadro dei manager e il riquadro dei progetti. In questo modo l'area di disegno viene ingrandita e consente di utilizzare più stream o stream di grandi dimensioni.

In alternativa, nel menu Visualizza fare clic su Palette nodi, Manager o Progetto per attivare o disattivare la visualizzazione di questi elementi.

Figura 2-12
Area di disegno dello stream ingrandita



In alternativa alla chiusura della palette dei nodi e dei riquadri dei manager e dei progetti è possibile utilizzare l'area di disegno come pagina scorrevole spostandosi in verticale e in orizzontale tramite le barre di scorrimento poste di fianco e in fondo alla finestra di SPSS Modeler.

È possibile anche controllare la visualizzazione del markup, costituito dai commenti agli stream, dai collegamenti dei modelli e dall'indicazione dei rami di calcolo del punteggio. Per attivare o disattivare la visualizzazione, fare clic su:

Visualizza > Markup stream

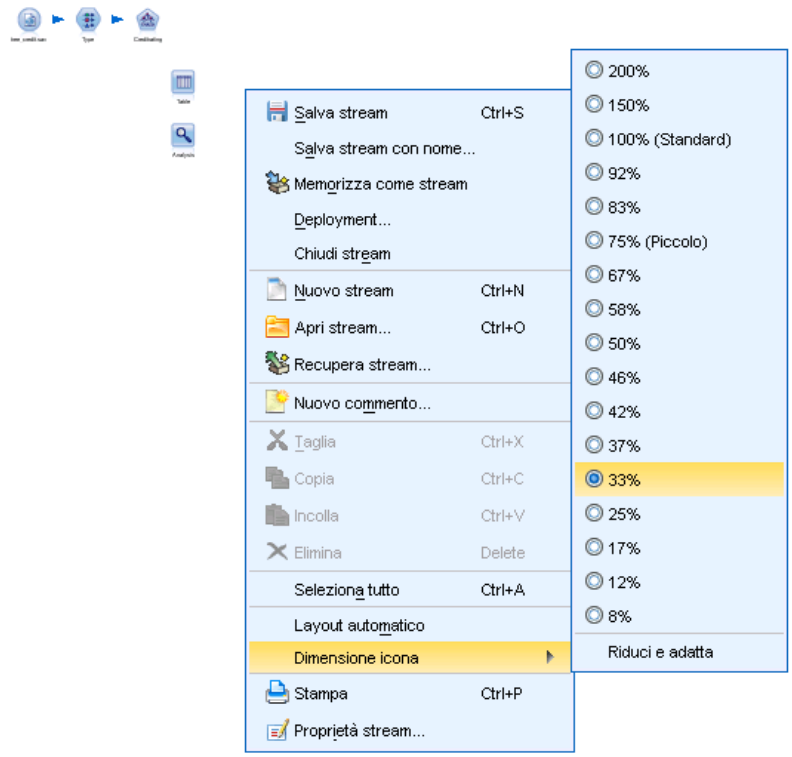
Modifica della dimensione dell'icona di uno stream

È possibile modificare le dimensioni delle icone degli stream nei modi seguenti.

- Attraverso l'impostazione di una proprietà dello stream
- Attraverso un menu di scelta rapida dello stream
- Utilizzando la tastiera

È possibile ridimensionare l'intera visualizzazione dello stream su una scala compresa tra l'8% e il 200% della dimensione dell'icona standard.

Figura 2-13
Modifica della dimensione dell'icona



Per ridimensionare l'intero stream (attraverso le proprietà dello stream)

- Dal menu principale, scegliere
  Strumenti > Proprietà stream > Opzioni > Layout.
- Scegliere la dimensione desiderata dal menu Dimensione icona.
- Fare clic su Applica per visualizzare il risultato.
- Fare clic su OK per salvare la modifica.

Per ridimensionare l'intero stream (attraverso il menu)

- Fare clic con il pulsante destro del mouse sullo sfondo dello stream nell'area di disegno.
- Scegliere Dimensione icona e selezionare la dimensione desiderata.

Per ridimensionare l'intero stream (attraverso la tastiera)

- Premere Ctrl + [-] sulla tastiera principale per eseguire lo zoom indietro alla successiva dimensione più piccola.
- Premere Ctrl + Maiusc + [+] sulla tastiera principale per eseguire lo zoom avanti alla successiva dimensione più grande.

Questa funzione risulta particolarmente utile per ottenere una vista generale di uno stream complesso. La funzione consente inoltre di ridurre al minimo il numero di pagine necessario per la stampa di uno stream.

Utilizzo del mouse in IBM SPSS Modeler

Di seguito sono riportati gli utilizzi più comuni del mouse in IBM® SPSS® Modeler:

- **Clic singolo.** Utilizzare il pulsante destro o sinistro del mouse per scegliere opzioni dai menu, aprire menu di scelta rapida e accedere a numerosi altri controlli e opzioni standard. Fare clic e tenere premuto il pulsante per spostare e trascinare i nodi.
- **Doppio clic.** Fare doppio clic con il pulsante sinistro del mouse per posizionare i nodi nell'area di disegno dello stream e modificare i nodi esistenti.
- **Clic con il pulsante centrale.** Fare clic con il pulsante centrale del mouse e trascinare il cursore per connettere i nodi nell'area di disegno dello stream. Fare doppio clic con il pulsante centrale del mouse per disconnettere un nodo. Se non si dispone di un mouse con tre pulsanti, è possibile simulare questa funzionalità tenendo premuto Alt mentre si seleziona e si trascina l'elemento desiderato.

Utilizzo dei tasti di scelta rapida

In IBM® SPSS® Modeler, numerose operazioni di programmazione visuale sono associate a tasti di scelta rapida. Per esempio, per eliminare un nodo è possibile fare clic su di esso e premere il tasto Canc. Analogamente, per salvare rapidamente uno stream tenere premuto il tasto Ctrl e

premere contemporaneamente il tasto S. I comandi di controllo di questo tipo sono indicati da una combinazione di Ctrl più un altro tasto, per esempio Ctrl+S.

Nelle operazioni standard di Windows sono utilizzati diversi tasti di scelta rapida, per esempio Ctrl+X per tagliare. Questi tasti di scelta rapida sono supportati in SPSS Modeler insieme ai tasti di scelta rapida seguenti specifici per l'applicazione.

Nota: in alcuni casi, i precedenti tasti di scelta rapida utilizzati in SPSS Modeler sono in conflitto con quelli standard di Windows. I precedenti tasti di scelta rapida sono supportati con l'aggiunta del tasto Alt. Per esempio, Ctrl+Alt+C può essere utilizzato per attivare/disattivare la cache.

Tabella 2-1
Tasti di scelta rapida supportati

Tasto di scelta rapida	**Funzione**
Ctrl+A	Seleziona tutto
Ctrl+X	Taglia
Ctrl+N	Nuovo stream
Ctrl+O	Apri stream
Ctrl+P	Stampa
Ctrl+C	Copia
Ctrl+V	Incolla
Ctrl+Z	Annulla
Ctrl+Q	Seleziona tutti i nodi a valle del nodo selezionato
Ctrl+W	Deseleziona tutti i nodi a valle del nodo selezionato (si alterna con Ctrl+Q)
Ctrl+E	Esegui a partire dal nodo selezionato
Ctrl+S	Salva stream corrente
Alt+tasti freccia	Sposta i nodi selezionati nell'area di disegno dello stream in direzione della freccia utilizzata
Maiusc+F10	Apre il menu di scelta rapida per il nodo selezionato

Tabella 2-2
Scelte rapide supportate per i precedenti tasti di scelta rapida

Tasto di scelta rapida	**Funzione**
Ctrl+Alt+D	Duplica il nodo
Ctrl+Alt+L	Carica nodo
Ctrl+Alt+R	Rinomina il nodo
Ctrl+Alt+U	Crea il nodo input utente
Ctrl+Alt+C	Attiva/disattiva la cache
Ctrl+Alt+F	Svuota cache
Ctrl+Alt+X	Espandi Supernodo
Ctrl+Alt+Z	Zoom avanti/indietro
Canc	Elimina il nodo o la connessione

Stampa

In IBM® SPSS® Modeler è possibile stampare i seguenti oggetti:

- Diagrammi di stream
- Grafici
- Tabelle
- Report (dal nodo Report e dai report del progetto)
- Script (dalle finestre di dialogo Proprietà stream, Script locale o Script Supernodo)
- Modelli (browser Modello, schede di finestre di dialogo con stato attivo corrente, visualizzatori albero)
- Annotazioni (utilizzando la scheda Annotazioni per l'output)

Per stampare un oggetto:

- Per stampare senza anteprima, fare clic sul pulsante Stampa sulla barra degli strumenti.
- Per impostare la pagina prima della stampa, scegliere Imposta pagina dal menu File.
- Per richiamare l'anteprima prima della stampa, scegliere Anteprima di stampa dal menu File.
- Per visualizzare la finestra di dialogo di stampa standard con le opzioni di selezione delle stampanti e specificare le opzioni relative all'aspetto, scegliere Stampa dal menu File.

Automatizzazione di IBM SPSS Modeler

Il data mining può essere un processo complesso e talvolta lungo. Per questo motivo IBM® SPSS® Modeler comprende diversi tipi di supporto per la codifica e l'automazione.

- **Control Language for Expression Manipulation** (CLEM) è un linguaggio per l'analisi e la manipolazione dei flussi di dati negli stream di SPSS Modeler. I data miner fanno ampio ricorso al linguaggio CLEM nelle operazioni stream per eseguire attività sia semplici, come la derivazione dei profitti dai dati relativi ai costi e alle entrate, sia complesse, come la trasformazione dei dati dei registri Web in un insieme di campi e record contenenti informazioni utilizzabili. Per ulteriori informazioni, vedere l'argomento Informazioni su CLEM in il capitolo 7 in *Manuale dell'utente di IBM SPSS Modeler 15*.
- **Gli script** sono un potente strumento per automatizzare i processi nell'interfaccia utente. Possono eseguire lo stesso tipo di azioni compiute dagli utenti con un mouse o una tastiera. È possibile impostare le opzioni per i nodi ed eseguire creazioni di campi utilizzando un sottoinsieme di CLEM. Inoltre è possibile specificare l'output e manipolare i modelli generati. Per ulteriori informazioni, vedere l'argomento Panoramica sugli script in il capitolo 2 in *IBM SPSS Modeler 15 Guida per script e automazione*.

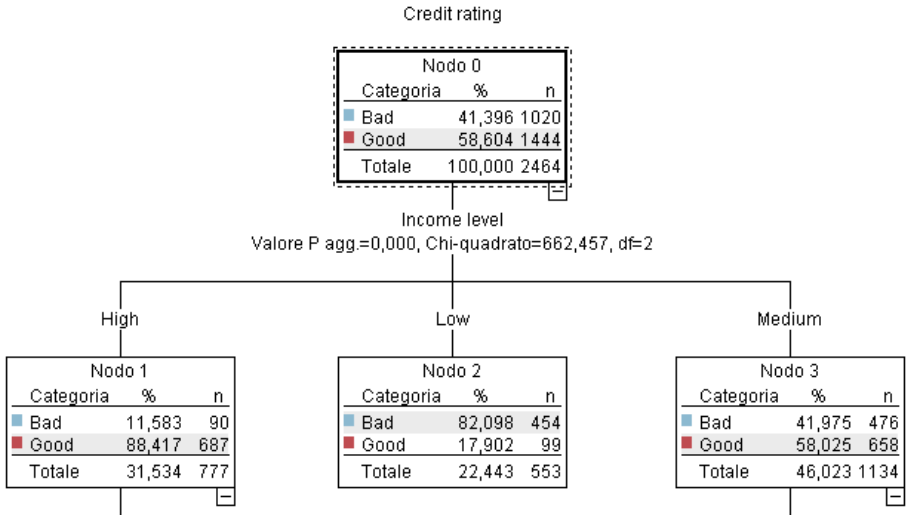
Capitolo
3

Introduzione alla modellazione

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

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

Figura 3-1
Modello di albero decisionale semplice



In questo esempio viene utilizzato un modello **Albero decisionale** che classifica i record (e prevede una risposta) mediante una serie di regole decisionali, per esempio:

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

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

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

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

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

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

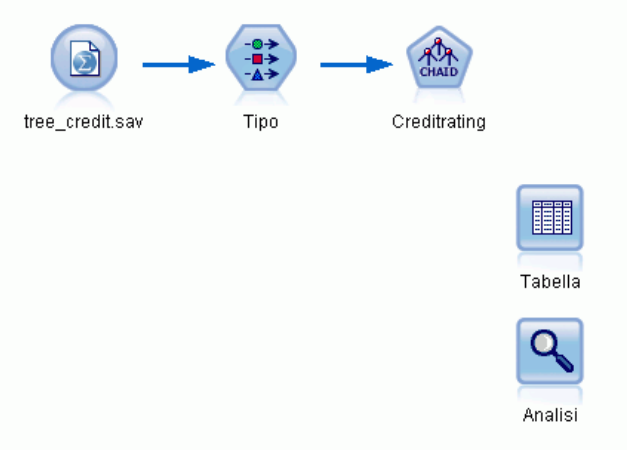
In questo esempio viene utilizzato lo stream denominato *modelingintro.str*, disponibile nella sottocartella *streams* della cartella *Demos*. Il file di dati è *tree_credit.sav*. Per ulteriori informazioni, vedere l'argomento Cartella Demos in il capitolo 1 in *Manuale dell'utente di IBM SPSS Modeler 15*.

Si osservi lo stream.

- Dal menu principale scegliere:
  File > Apri stream
- Fare clic sull'icona della pepita d'oro nella barra degli strumenti della finestra di dialogo Apri e scegliere la cartella Demos.
- Fare doppio clic sulla cartella *streams*.
- Fare doppio clic sul file denominato *modelingintro.str*.

Creazione dello stream

Figura 3-2
Stream di modellazione



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

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

In questo esempio verrà utilizzato un nodo Modelli CHAID. CHAID, o Chi-squared Automatic Interaction Detection, è un metodo di classificazione che crea alberi decisionali utilizzando un particolare tipo di statistica, noto come statistica chi-quadrato, per calcolare i punti migliori in cui eseguire le suddivisioni nell'albero decisionale.

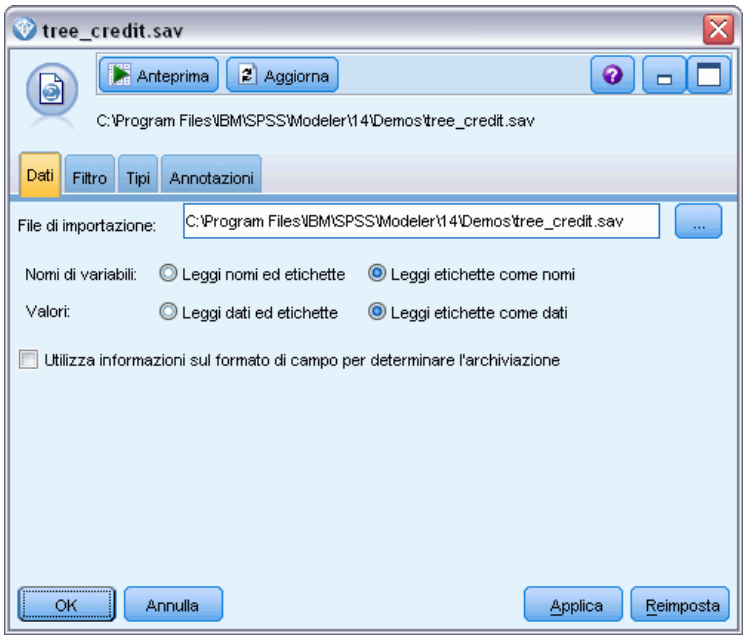
Se il nodo di input specifica già i livelli di misurazione, è possibile eliminare il nodo Tipo. Dal punto di vista funzionale, il risultato è identico.

Lo stream è dotato inoltre di nodi Tabella e Analisi che saranno utilizzati per visualizzare i risultati del calcolo del punteggio dopo la creazione e l'aggiunta allo stream dell'insieme di modelli.

Il nodo di input File Statistics legge i dati in formato SPSS Statistics dal file di dati *tree_credit.sav*, installato nella cartella *Demos* (per fare riferimento a questa cartella nell'installazione corrente di IBM® SPSS® Modeler viene utilizzata una variabile speciale denominata *$CLEO_DEMOS*,

al fine di garantire che il percorso sia valido indipendentemente dalla cartella o dalla versione dell'installazione corrente).

Figura 3-3
Lettura di dati con un nodo di input File Statistics

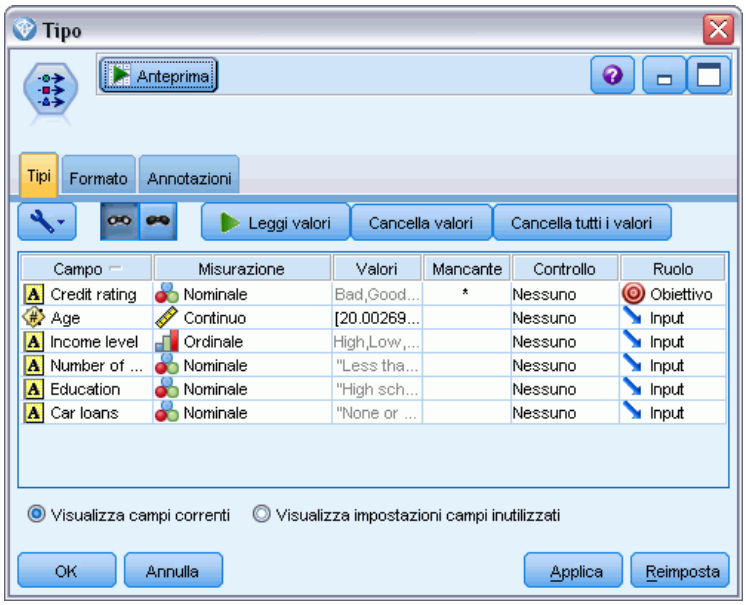


Il nodo Tipo specifica il **livello di misurazione** per ogni campo. Il livello di misurazione è una categoria che indica il tipo di dati all'interno del campo. Il file di dati di origine utilizzato in questo esempio impiega tre diversi livelli di misurazione.

Un campo **Continuo** (per esempio il campo *Età*) contiene valori numerici continui, mentre un campo **Nominale** (per esempio il campo *Rischio creditizio*) ha due o più valori distinti, per esempio *Sfavorevole*, *Favorevole* o *Nessuno storico crediti*. Un campo **Ordinale** (per esempio

il campo *Livello di reddito*) descrive dati con più valori distinti dotati di un ordine intrinseco, in questo caso *Basso*, *Medio* e *Alto*.

Figura 3-4
Impostazione dei campi obiettivo e di input con il nodo Tipo



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

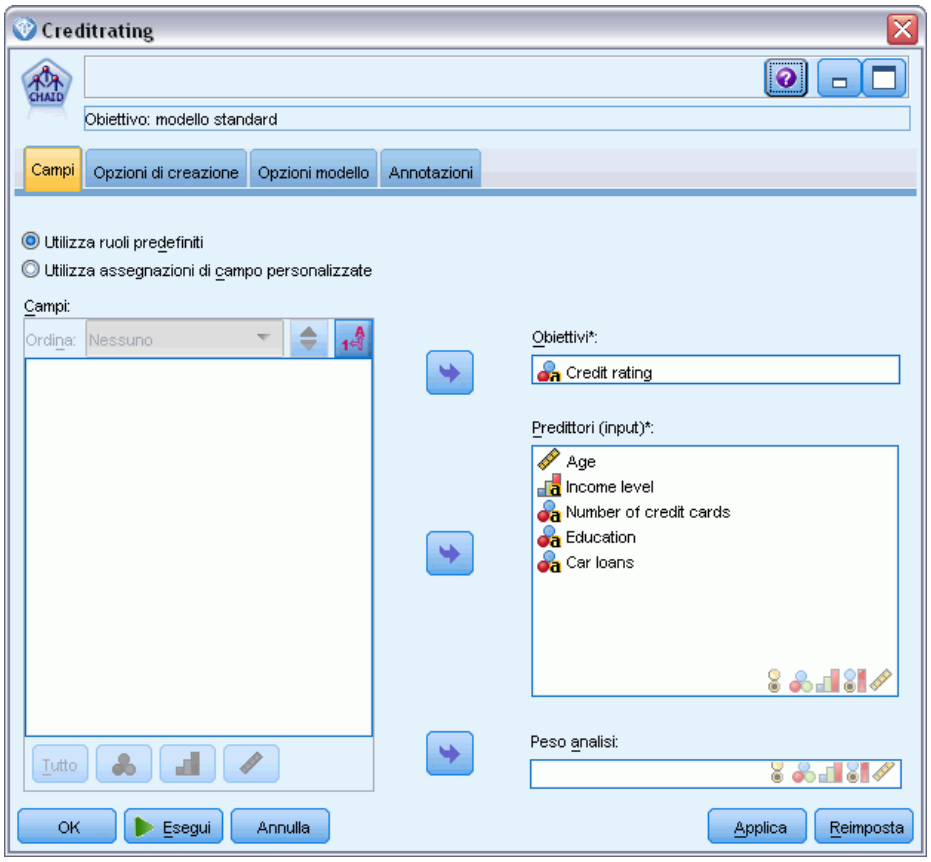
Per gli altri campi, il ruolo viene impostato su *Input*. I campi di input sono noti anche come **predittori**, o campi i cui valori sono utilizzati dall'algoritmo di modellazione per prevedere il valore del campo obiettivo.

Il nodo Modelli CHAID genera il modello.

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

▶ Fare clic sulla scheda Opzioni di creazione.

Figura 3-5
Nodo Modelli CHAID, scheda Campi



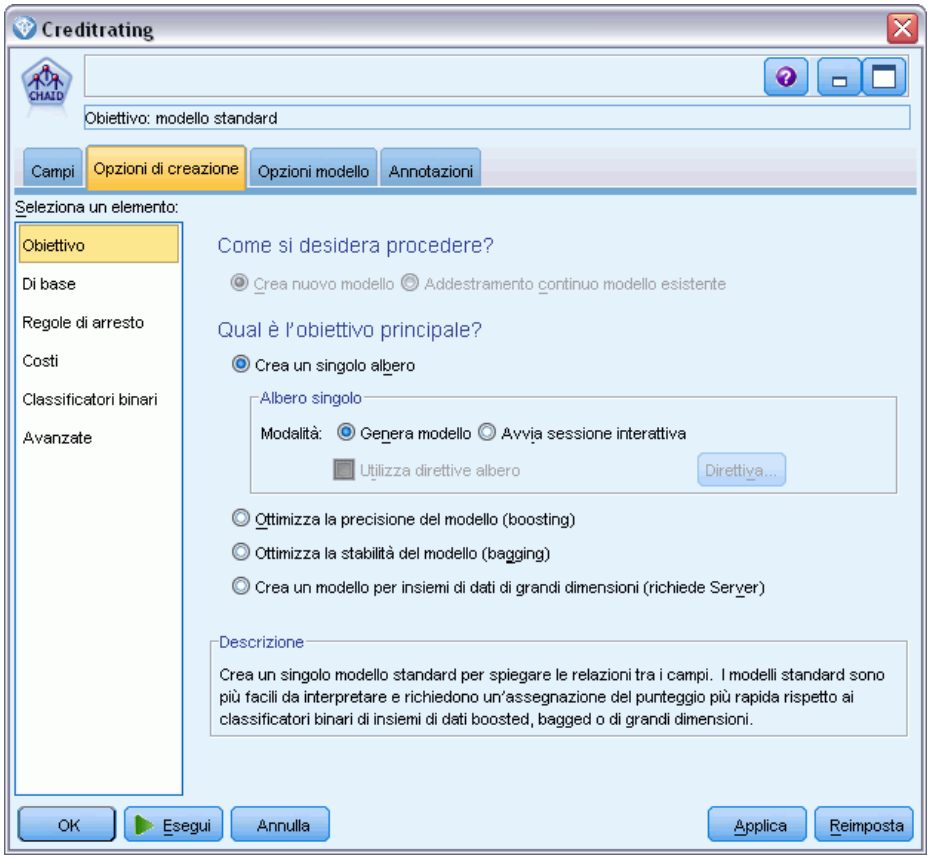
La scheda include diverse opzioni che consentono di specificare il tipo di modello che si desidera creare.

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

Poiché si desidera creare un solo modello di albero decisionale standard senza alcuna modifica, si lascia selezionata l'opzione di default Crea un singolo albero.

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

Figura 3-6
Nodo Modelli CHAID, scheda Opzioni di creazione

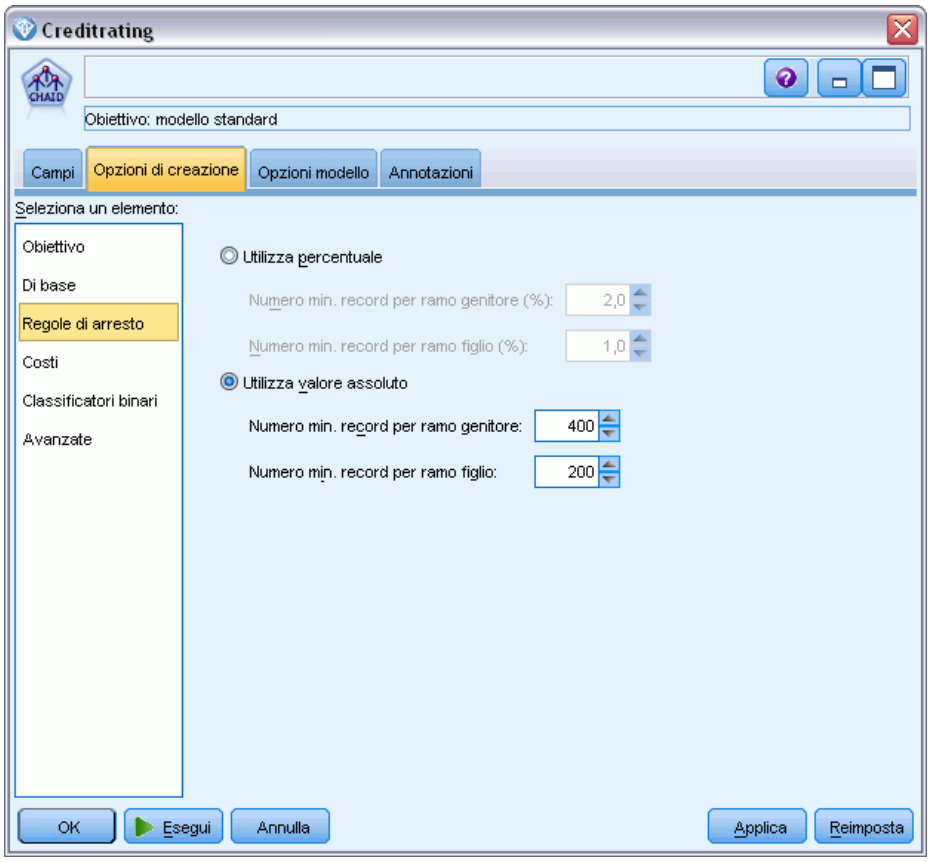


Per questo esempio l'albero dovrà essere abbastanza semplice, per cui se ne limiterà l'espansione aumentando il numero minimo di casi per i nodi genitore e figlio.

- Nella scheda Opzioni di creazione, selezionare Regole di arresto dal riquadro di navigazione di sinistra.
- Selezionare l'opzione Utilizza valore assoluto.
- Impostare Numero min. record per ramo genitore su 400.

▶ Impostare Numero min. record per ramo figlio su 200.

Figura 3-7
Impostazione dei criteri di arresto per la creazione di alberi decisionali



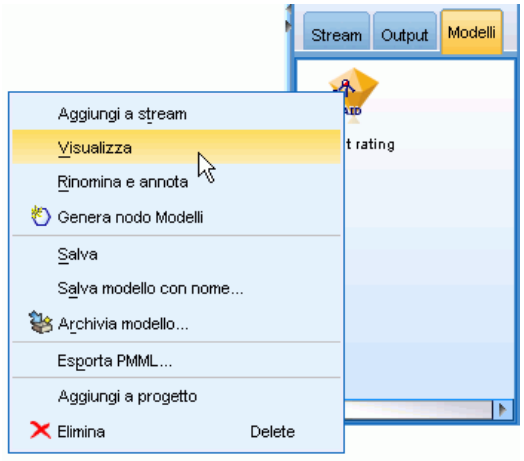
Poiché è possibile utilizzare tutte le opzioni di default nell'esempio, fare clic su Esegui per creare il modello. In alternativa, fare clic sul nodo con il pulsante destro del mouse e scegliere Esegui dal menu di scelta rapida oppure selezionare il nodo e scegliere Esegui dal menu Strumenti.

Visualizzazione del modello

Una volta terminata l'esecuzione, l'insieme di modelli viene aggiunto alla palette Modelli nell'angolo superiore destro della finestra dell'applicazione e anche all'area di disegno dello stream con un collegamento al nodo Modelli con cui è stato creato. Per visualizzare i dettagli del

modello, fare clic con il pulsante destro del mouse sull'insieme di modelli e scegliere Visualizza (dalla palette Modelli) o Modifica (dall'area di disegno).

Figura 3-8
Palette Modelli



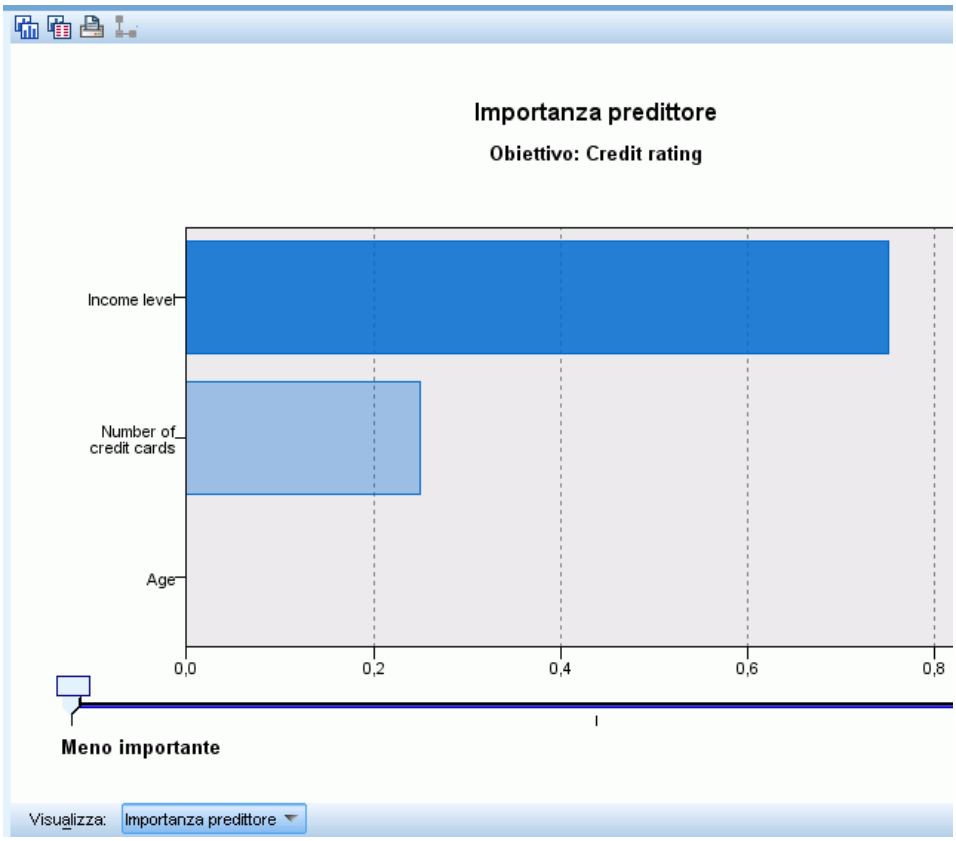
Nel caso dell'insieme di modelli CHAID, la scheda Modello visualizza i dettagli sotto forma di insieme di regole ovvero, in sostanza, una serie di regole che è possibile utilizzare per assegnare singoli record a nodi figlio in base ai valori dei vari campi di input.

Figura 3-9
Insieme di modelli CHAID, insieme di regole

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

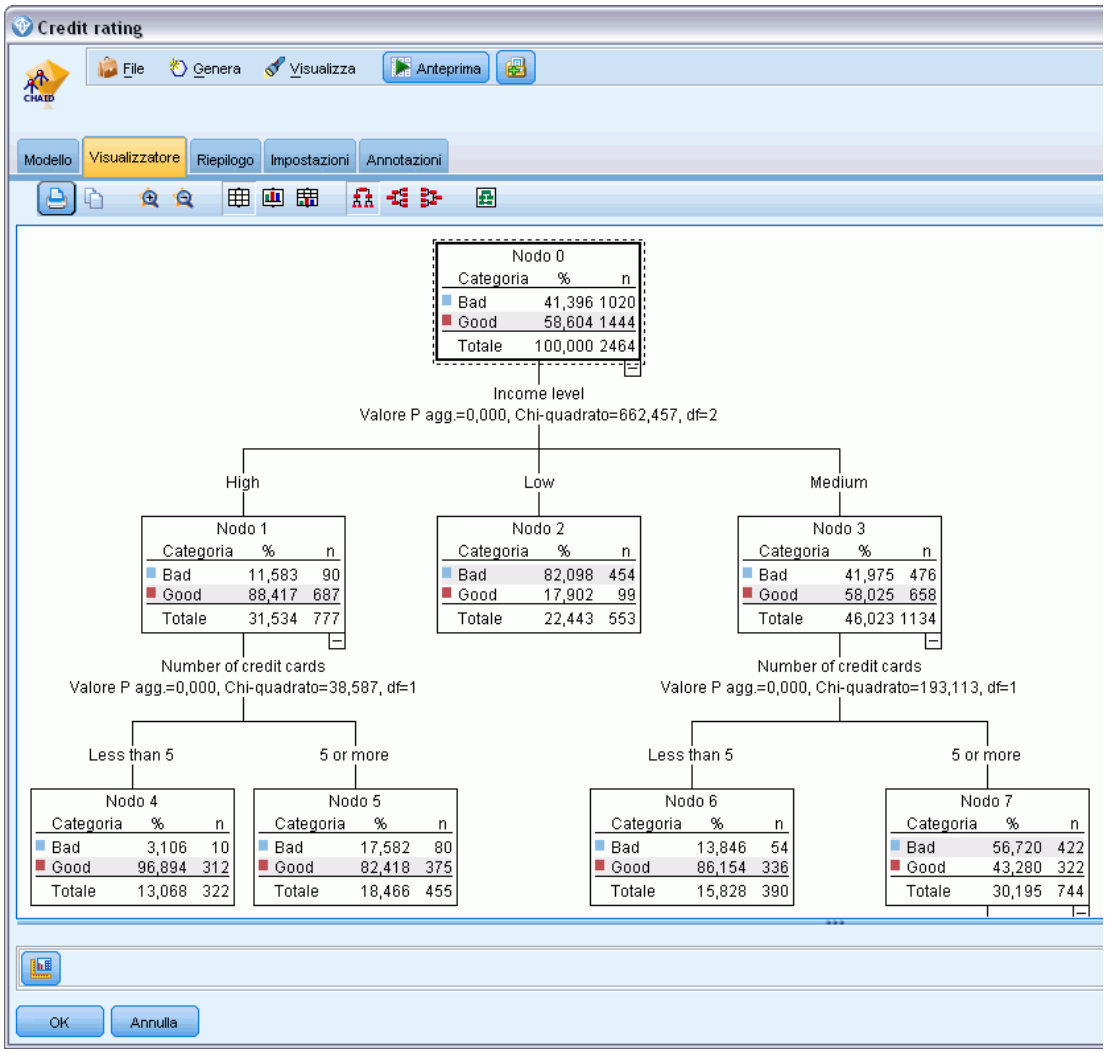
A destra dell'insieme di regole, la scheda Modello visualizza il grafico dell'importanza dei predittori, che visualizza l'importanza relativa di ciascun predittore nella stima del modello. Dal grafico si nota che *Livello di reddito* è probabilmente l'elemento più significativo in questo caso, e che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

Figura 3-10
Grafico dell'importanza dei predittori



La scheda Visualizzatore dell'insieme di modelli mostra lo stesso modello sotto forma di albero, con un nodo in corrispondenza di ogni punto decisionale. Per ingrandire un determinato nodo o ridurlo in modo da visualizzare meglio l'albero è possibile utilizzare i comandi di ingrandimento e riduzione.

Figura 3-11
Scheda Visualizzatore dell'insieme di modelli con il comando di riduzione selezionato



Osservando la parte superiore dell'albero, il primo nodo (il Nodo 0) fornisce un riepilogo di tutti i record dell'insieme di dati. Poco più del 40% dei casi dell'insieme di dati è classificato come a rischio creditizio sfavorevole. È una proporzione piuttosto elevata. Per capire da quali fattori potrebbe dipendere, è possibile osservare l'albero alla ricerca di indizi.

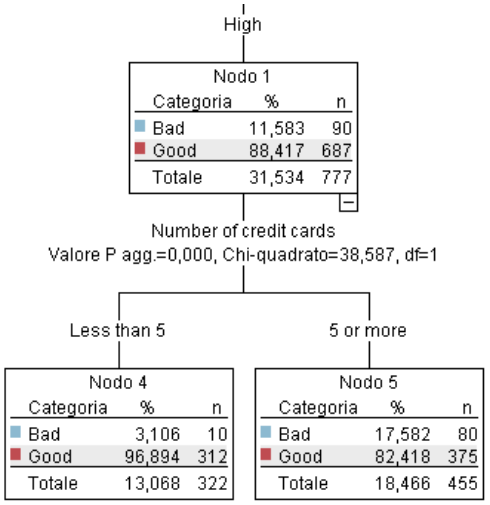
Si nota che la prima suddivisione è in corrispondenza di *Livello di reddito*. I record in cui il livello di reddito rientra nella categoria *Basso* vengono assegnati al Nodo 2, e non sorprende rilevare che questa categoria contiene la percentuale più elevata di clienti insolventi. Chiaramente, concedere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

Tuttavia, il 16% dei clienti di questa categoria in realtà *ha rimborsato* il prestito, per cui la previsione non sarà sempre corretta. Nessun modello è in grado di prevedere ogni singola risposta, ma un modello efficace dovrebbe consentire di prevedere la risposta *più probabile* per ogni record in base ai dati disponibili.

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

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

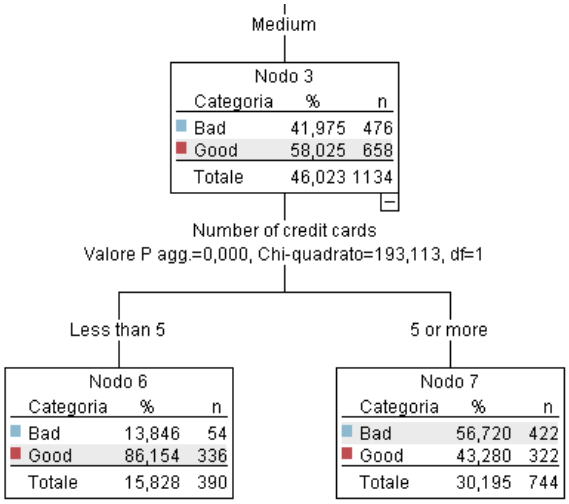
Figura 3-12
Visualizzazione struttura dei clienti con reddito elevato



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

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

Figura 3-13
Visualizzazione struttura dei clienti con reddito medio



Si è visto dunque che ogni record inserito in questo modello verrà assegnato a un nodo specifico e che gli verrà assegnata una previsione *Favorevole* o *Sfavorevole* a seconda della risposta più comune per quel nodo.

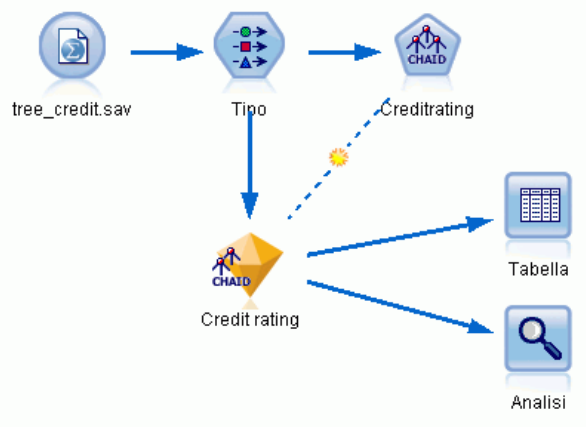
Questo processo di assegnazione delle previsioni ai singoli record è noto come **calcolo del punteggio**. Calcolando il punteggio degli stessi record utilizzati per la stima del modello è possibile valutare il grado di precisione delle relative previsioni sui dati di addestramento (i dati di cui conosciamo il risultato). Ecco come procedere.

Valutazione del modello

Per comprendere il funzionamento del calcolo del punteggio occorre visualizzare il modello. Per valutare invece il *grado di precisione* del modello è necessario calcolare il punteggio di alcuni record e confrontare le risposte previste dal modello con i risultati effettivi. In questo modo sarà

calcolato il punteggio degli stessi record utilizzati per la stima del modello, il che consentirà di confrontare le risposte osservate e quelle previste.

Figura 3-14
Collegamento dell'insieme di modelli ai nodi di output per la valutazione del modello



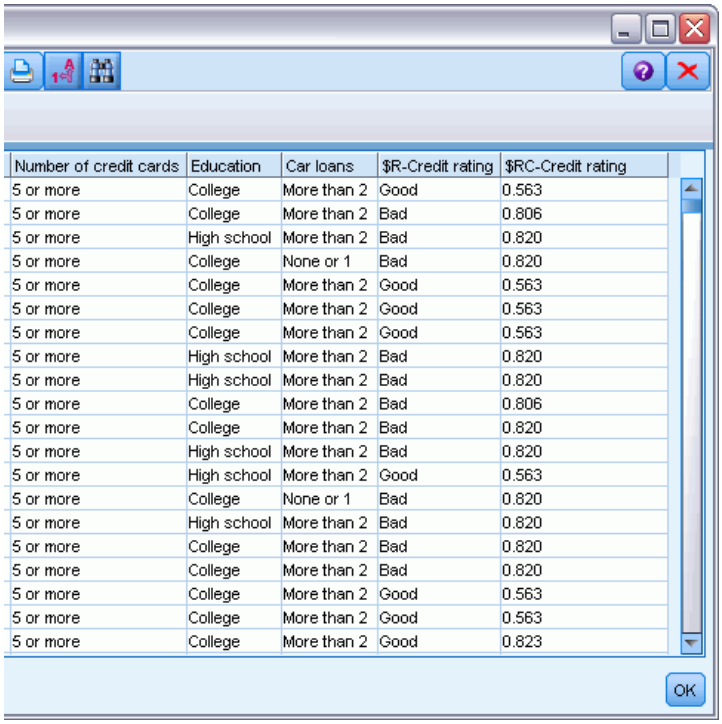
- Per visualizzare i punteggi o le previsioni, associare il nodo Tabella all'insieme di modelli, fare doppio clic sul nodo Tabella e scegliere Esegui.

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

  Per convenzione, i nomi dei campi generati durante il calcolo del punteggio hanno lo stesso nome del campo obiettivo, preceduto però da un prefisso standard quale *$R-* per le previsioni o *$RC-* per i valori di confidenza. I vari tipi di modelli adottano insiemi di prefissi diversi. Un **valore**

di confidenza è la stima effettuata dal modello, su una scala da 0,0 a 1,0, della precisione dei singoli valori previsti.

Figura 3-15
Tabella che mostra i punteggi generati e i valori di confidenza



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

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

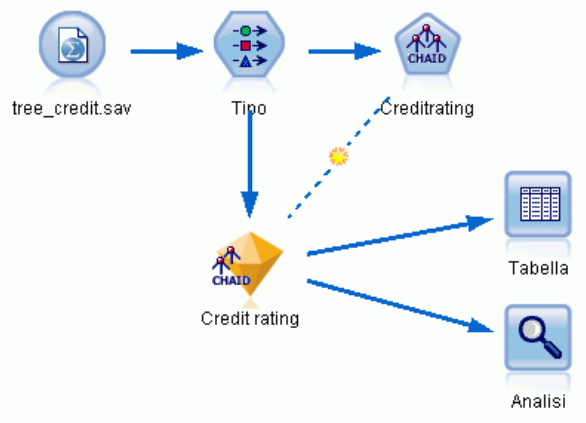
Per evitare questo inconveniente è possibile continuare a suddividere l'albero in rami sempre più piccoli, fino a che ogni nodo risulti al 100% —tutto *Favorevole* o *Sfavorevole*, senza risposte miste. Un modello del genere sarebbe però molto complicato e probabilmente non verrebbe correttamente generalizzato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

Per determinare con esattezza il numero delle previsioni corrette si potrebbe scorrere la tabella e contare i record in cui il valore del campo previsto *$R-Rischio creditizio* corrisponde al valore di *Rischio creditizio*. Per fortuna esiste un metodo molto più semplice: si può utilizzare un nodo Analisi, che esegue questa operazione automaticamente.

- Collegare l'insieme di modelli al nodo Analisi.

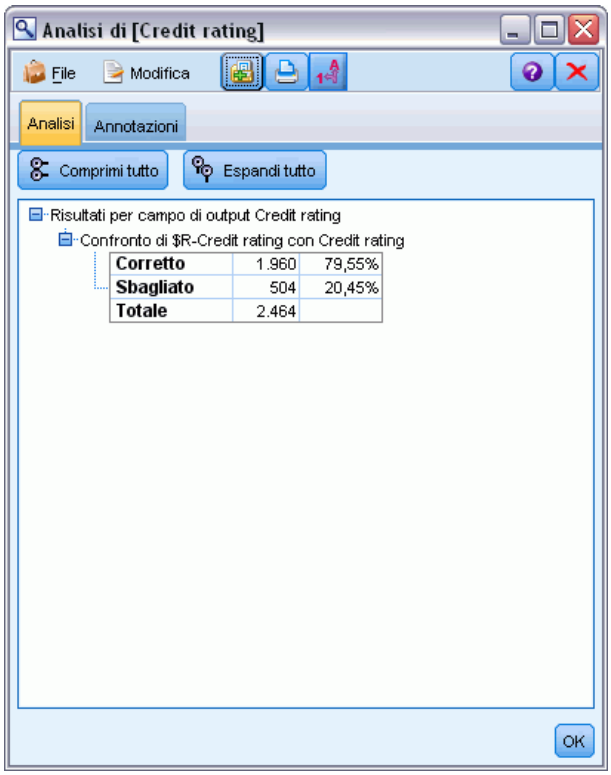
▶ Fare doppio clic sul nodo Analisi e fare clic su Esegui.

Figura 3-16
Collegamento di un nodo Analisi



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

Figura 3-17
Risultati dell'analisi da confronto tra risposte previste e osservate



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

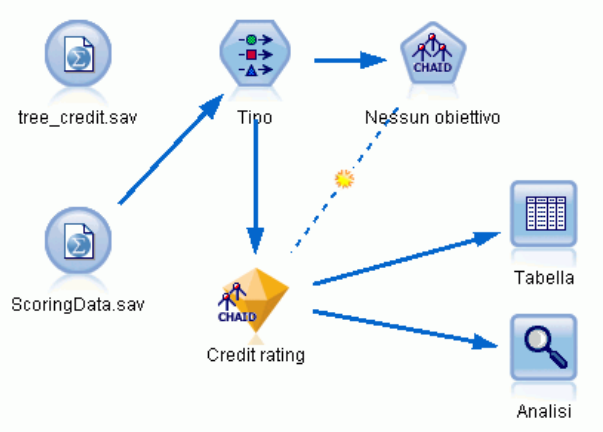
Se si utilizza una partizione campione per generare il modello e un altro campione per verificarlo è possibile ottenere un'indicazione molto più precisa della facilità di generalizzazione su altri insiemi di dati.

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

Calcolo del punteggio dei record

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

Figura 3-18
Collegamento di nuovi dati per il calcolo del punteggio



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

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

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

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

Riepilogo

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

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

Capitolo
4

Modellazione automatica per un obiettivo flag

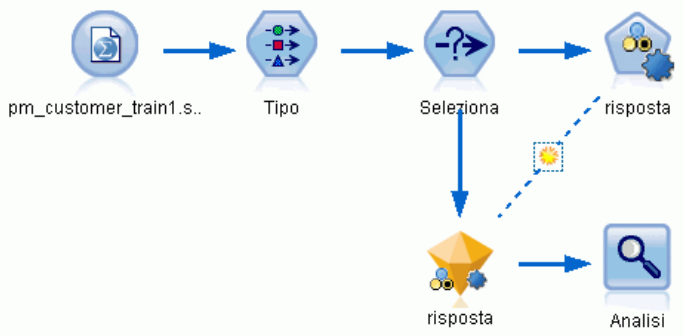
Modellazione della risposta dei clienti (Classificatore automatico)

Il nodo Classificatore automatico consente di creare e confrontare automaticamente una serie di modelli diversi per risultati flag (per esempio, quelli volti a prevedere se un determinato cliente restituirà o meno un prestito o risponderà a una determinata offerta) o per obiettivi nominali (insieme). In questo esempio, si cerca un risultato flag (sì o no). Con uno stream relativamente semplice, il nodo genera e classifica un insieme di modelli candidati, sceglie quelli con le migliori prestazioni e li combina in un unico modello aggregato (generato mediante un nodo Classificatore binario). Questo approccio unisce la facilità di automazione al vantaggio di combinare più modelli, che spesso generano previsioni più accurate di quelle che si otterrebbero con un solo modello.

Questo esempio è basato su una società fittizia che desidera ottenere risultati più redditizi inviando offerte mirate ai singoli clienti.

L'approccio adottato sfrutta i vantaggi dell'automazione. Per un esempio simile che utilizza un obiettivo continuo (intervallo numerico), vedere il capitolo 5 a pag. 57.

Figura 4-1
Stream di esempio di Classificatore automatico



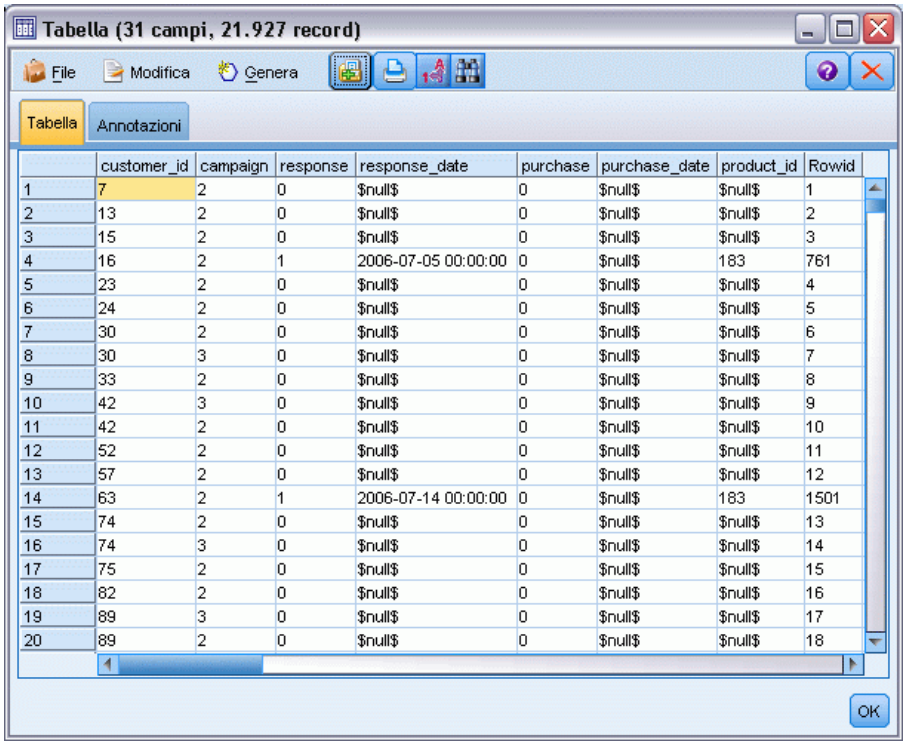
Questo esempio utilizza lo stream *pm_binaryclassifier.str*, installato nella cartella Demos in *streams*. Il file di dati utilizzato è *pm_customer_train1.sav*. Per ulteriori informazioni, vedere l'argomento Cartella Demos in il capitolo 1 a pag. 7.

Dati storici

Il file *pm_customer_train1.sav* include i dati storici che tengono traccia delle offerte fatte a specifici clienti nell'ambito delle campagne precedenti, come indicato dal campo *campagna*. Il numero maggiore di record rientra nella campagna *Premium account*.

I valori del campo *campagna* vengono codificati come numeri interi nei dati (per esempio, *2* = *Premium account*). In seguito si definiranno delle etichette per questi valori da utilizzare per ottenere risultati più significativi.

Figura 4-2
Dati sulle promozioni precedenti



	customer_id	campaign	response	response_date	purchase	purchase_date	product_id	Rowid
1	7	2	0	$null$	0	$null$	$null$	1
2	13	2	0	$null$	0	$null$	$null$	2
3	15	2	0	$null$	0	$null$	$null$	3
4	16	2	1	2006-07-05 00:00:00	0	$null$	183	761
5	23	2	0	$null$	0	$null$	$null$	4
6	24	2	0	$null$	0	$null$	$null$	5
7	30	2	0	$null$	0	$null$	$null$	6
8	30	3	0	$null$	0	$null$	$null$	7
9	33	2	0	$null$	0	$null$	$null$	8
10	42	3	0	$null$	0	$null$	$null$	9
11	42	2	0	$null$	0	$null$	$null$	10
12	52	2	0	$null$	0	$null$	$null$	11
13	57	2	0	$null$	0	$null$	$null$	12
14	63	2	1	2006-07-14 00:00:00	0	$null$	183	1501
15	74	2	0	$null$	0	$null$	$null$	13
16	74	3	0	$null$	0	$null$	$null$	14
17	75	2	0	$null$	0	$null$	$null$	15
18	82	2	0	$null$	0	$null$	$null$	16
19	89	3	0	$null$	0	$null$	$null$	17
20	89	2	0	$null$	0	$null$	$null$	18

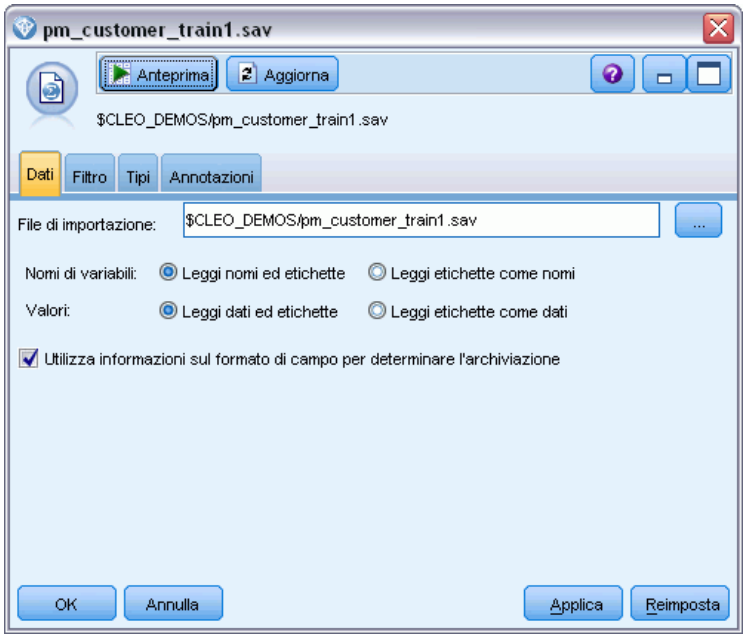
Il file include inoltre un campo *risposta* che indica se l'offerta è stata accettata (0 = *no* e 1 = *sì*). Questo è il **campo obiettivo** o il valore che si vuole prevedere. Il file include inoltre un numero di campi contenenti informazioni finanziarie e demografiche relative a ciascun cliente, che possono essere utilizzate per generare o "addestrare" un modello che prevede i tassi di risposta relativi a singoli o gruppi di clienti diversi in base a caratteristiche specifiche, quali reddito, età o numero di transazioni per mese.

Creazione dello stream

- Aggiungere un nodo di input File Statistics che punti al file *pm_customer_train1.sav* situato nella directory *Demos* dell'installazione IBM® SPSS® Modeler. (In alternativa, è possibile specificare

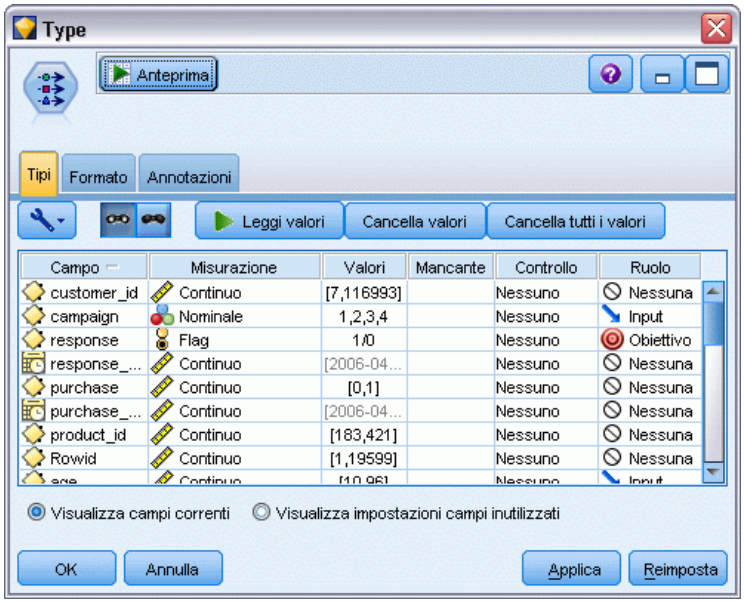
$CLEO_DEMOS/ nel percorso del file come collegamento per fare riferimento a questa cartella. Si noti che nel percorso è necessario utilizzare la barra e non la barra rovesciata, come mostrato.)

Figura 4-3
Lettura dei dati



- Aggiungere un nodo Tipo e selezionare *risposta* come campo obiettivo (Ruolo = Obiettivo). Impostare la Misurazione di questo campo su Flag.

Figura 4-4
Impostazione del livello di misurazione e del ruolo

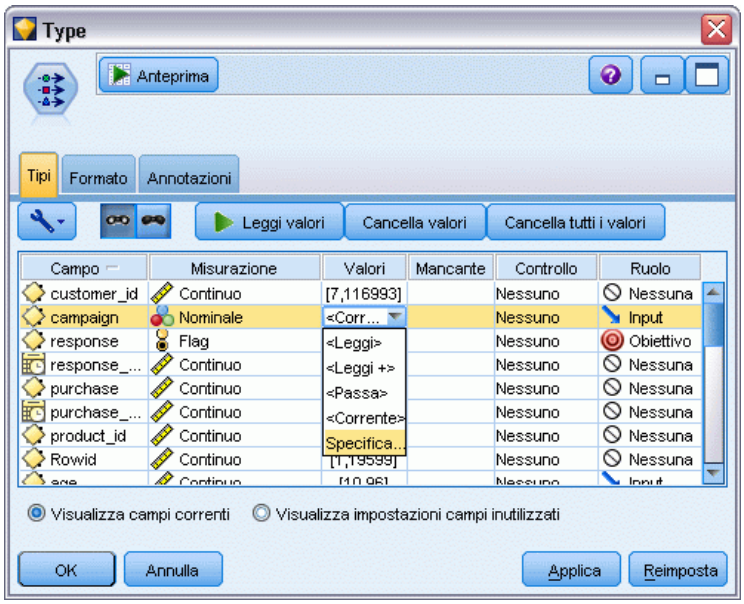


- Impostare su Nessuno il ruolo dei seguenti campi: *id_cliente*, *campagna*, *data_risposta*, *acquisto*, *data_acquisto*, *id_prodotto*, *IDriga* e *X_casuale*. Alla creazione del modello, questi campi verranno ignorati.

- Fare clic sul pulsante Leggi valori nel nodo Tipo per assicurarsi che tali valori vengano istanziati.

  Come visto in precedenza, i dati di origine comprendono informazioni relative a quattro diverse campagne, ciascuna mirata a un diverso tipo di account cliente. Nei dati, queste campagne sono codificate come numeri interi. Pertanto, per ricordare più agevolmente quale tipo di account rappresenta ciascun numero intero, definire delle etichette per ciascuno di essi.

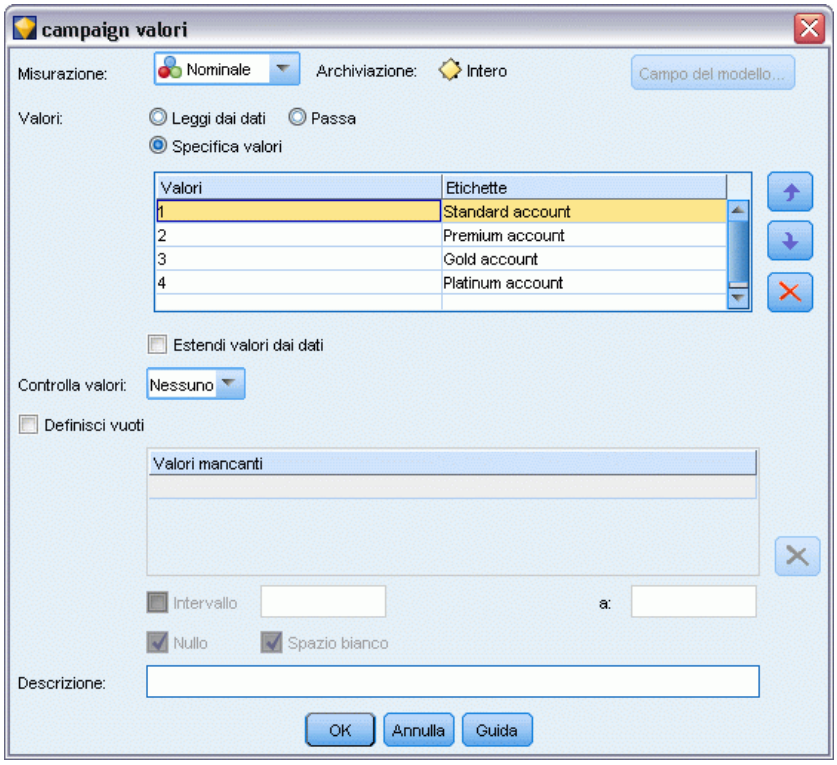
Figura 4-5
Scelta di specificare i valori per un campo



- Sulla riga relativa al campo campagna, fare clic sulla voce nella colonna Valori.

- Scegliere Specifica dall'elenco a discesa.

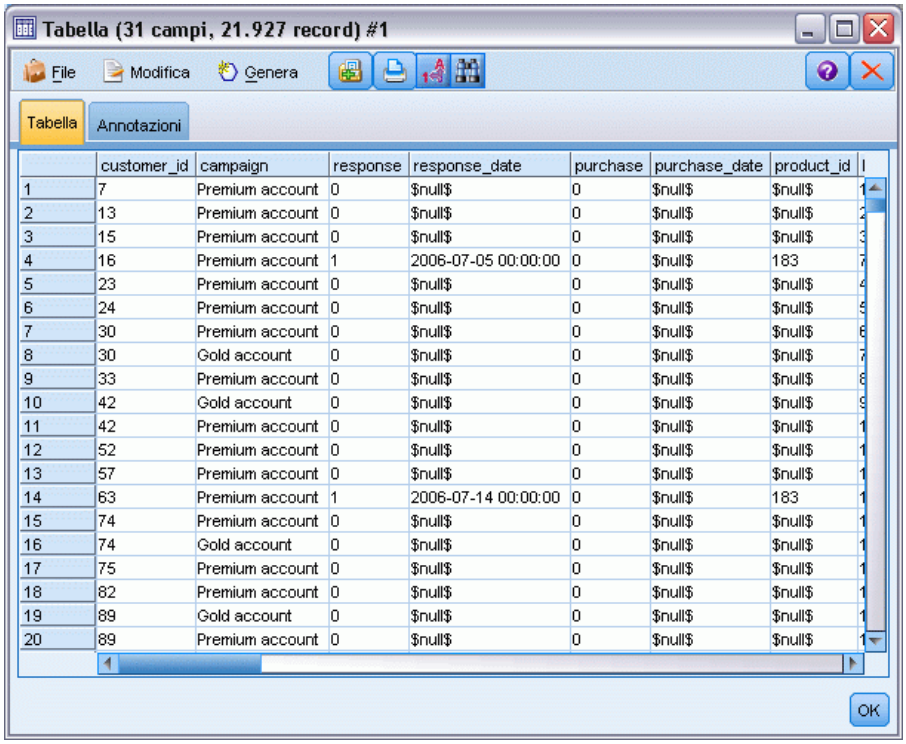
Figura 4-6
Definizione delle etichette per i valori dei campi



- Nella colonna Etichette, digitare le etichette come illustrato per ciascuno dei quattro valori del campo campagna.
- Fare clic su OK.

Ora è possibile visualizzare le etichette anziché i numeri interi nelle finestre di output.

Figura 4-7
Visualizzazione delle etichette dei valori dei campi

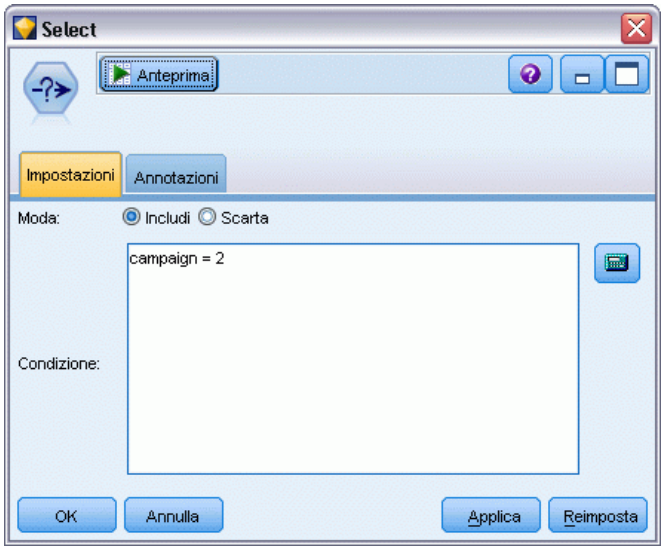


	customer_id	campaign	response	response_date	purchase	purchase_date	product_id	l
1	7	Premium account	0	$null$	0	$null$	$null$	1
2	13	Premium account	0	$null$	0	$null$	$null$	2
3	15	Premium account	0	$null$	0	$null$	$null$	3
4	16	Premium account	1	2006-07-05 00:00:00	0	$null$	183	7
5	23	Premium account	0	$null$	0	$null$	$null$	4
6	24	Premium account	0	$null$	0	$null$	$null$	5
7	30	Premium account	0	$null$	0	$null$	$null$	6
8	30	Gold account	0	$null$	0	$null$	$null$	7
9	33	Premium account	0	$null$	0	$null$	$null$	8
10	42	Gold account	0	$null$	0	$null$	$null$	9
11	42	Premium account	0	$null$	0	$null$	$null$	1
12	52	Premium account	0	$null$	0	$null$	$null$	1
13	57	Premium account	0	$null$	0	$null$	$null$	1
14	63	Premium account	1	2006-07-14 00:00:00	0	$null$	183	1
15	74	Premium account	0	$null$	0	$null$	$null$	1
16	74	Gold account	0	$null$	0	$null$	$null$	1
17	75	Premium account	0	$null$	0	$null$	$null$	1
18	82	Premium account	0	$null$	0	$null$	$null$	1
19	89	Gold account	0	$null$	0	$null$	$null$	1
20	89	Premium account	0	$null$	0	$null$	$null$	1

- Collegare un nodo Tabella al nodo Tipo.
- Aprire il nodo Tabella e fare clic su Esegui.
- Nella finestra di output, fare clic sul pulsante Visualizza etichette di valori e campi nella barra degli strumenti per visualizzare le etichette.
- Fare clic su OK per chiudere la finestra di output.

Sebbene i dati includano informazioni su quattro diverse campagne, l'analisi verrà focalizzata su una campagna alla volta. Poiché il numero maggiore di record rientra nella campagna Premium account (codificata come *campagna=2* nei dati), è possibile utilizzare il nodo Seleziona per includere nello stream solo i suddetti record.

Figura 4-8
Selezione di record per un'unica campagna

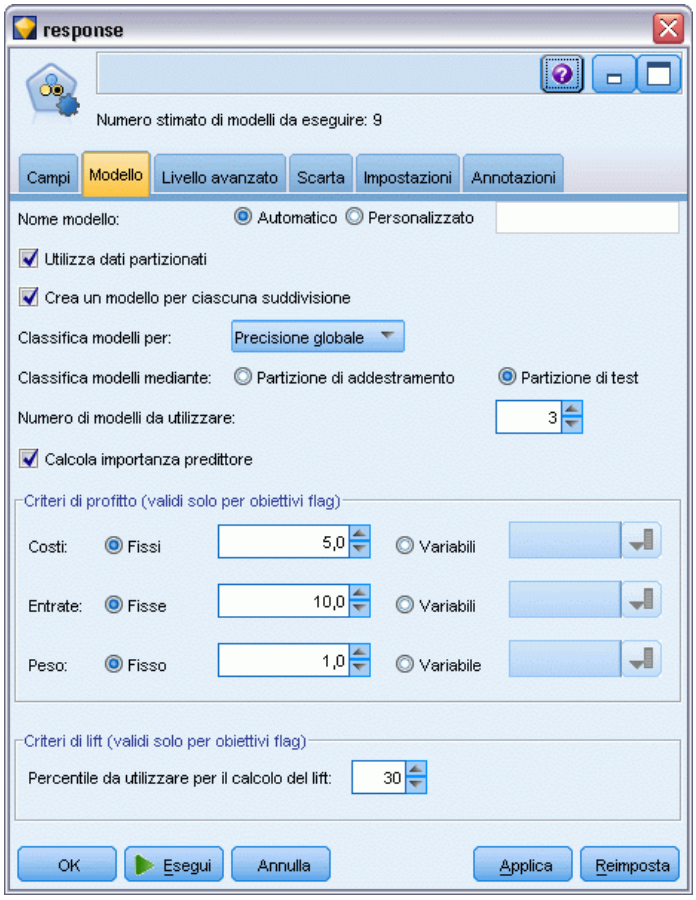


Generazione e confronto dei modelli

- Collegare un nodo Classificatore automatico e selezionare Precisione globale come misura per classificare i modelli.

- Impostare Numero di modelli da utilizzare su 3. Questo significa che quando si esegue il nodo vengono creati i tre modelli migliori.

Figura 4-9
Scheda Modello del nodo Classificatore automatico

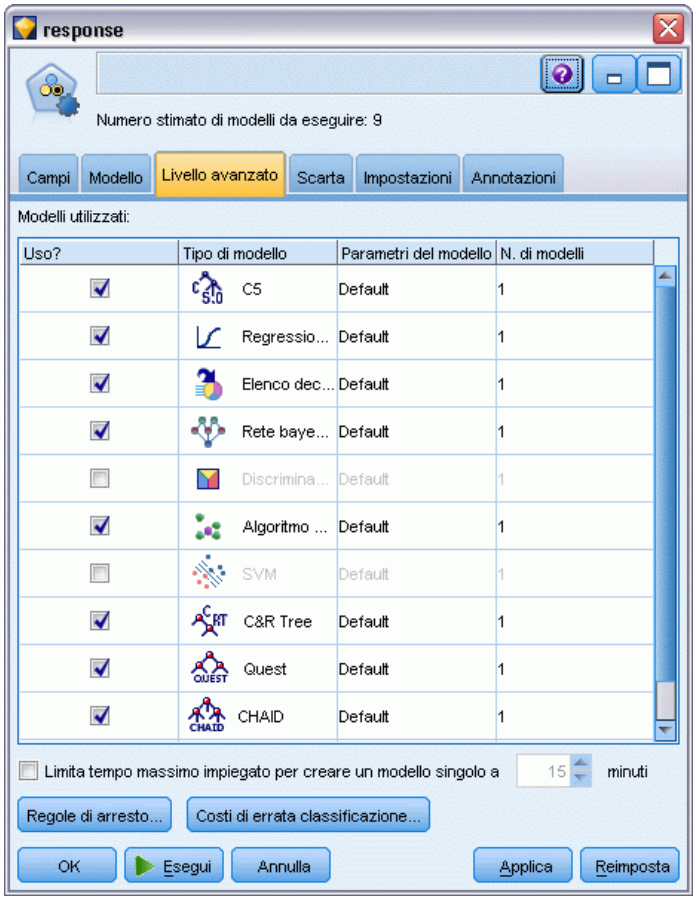


Nella scheda Livello avanzato è possibile scegliere tra un massimo di 11 diversi algoritmi di creazione dei modelli.

- Deselezionare i tipi di modello Discriminante e SVM. L'addestramento di questi modelli con i dati disponibili richiede più tempo rispetto agli altri. Deselezionandoli, l'esempio risulterà più rapido. Se invece non dispiace attendere, lasciarli pure selezionati.

Poiché il Numero di modelli da utilizzare nella scheda Modelli è stato impostato su 3, il nodo calcola la precisione dei nove algoritmi restanti e crea un unico insieme di modelli contenente i tre più accurati.

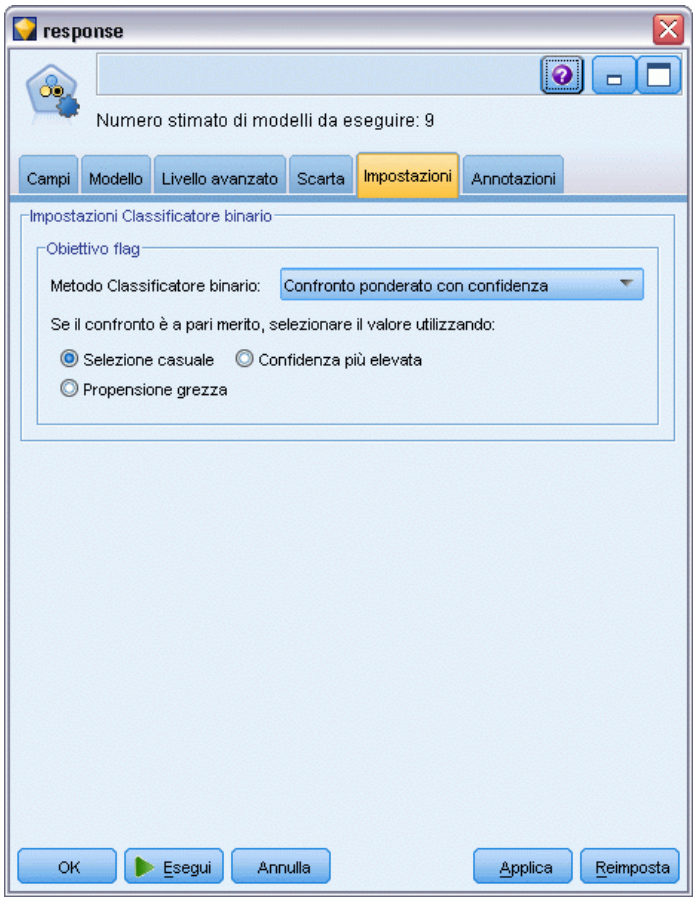
Figura 4-10
Scheda Livello avanzato del nodo Classificatore automatico



- Nella scheda Impostazioni, per il Metodo Classificatore binario, selezionare Confronto ponderato con confidenza. Questa opzione stabilisce come viene prodotto un singolo punteggio aggregato per ogni record.

Con un confronto semplice, se due modelli su tre prevedono *sì*, allora il *sì* vince per un voto di 2 a 1. Nel caso del confronto ponderato con confidenza, i voti sono ponderati in base al valore di confidenza di ogni previsione. Pertanto, se un modello prevede *no* con una confidenza maggiore delle due previsioni *sì* combinate, allora vince il *no*.

Figura 4-11
Nodo Classificatore automatico: scheda Impostazioni



► Fare clic su Esegui.

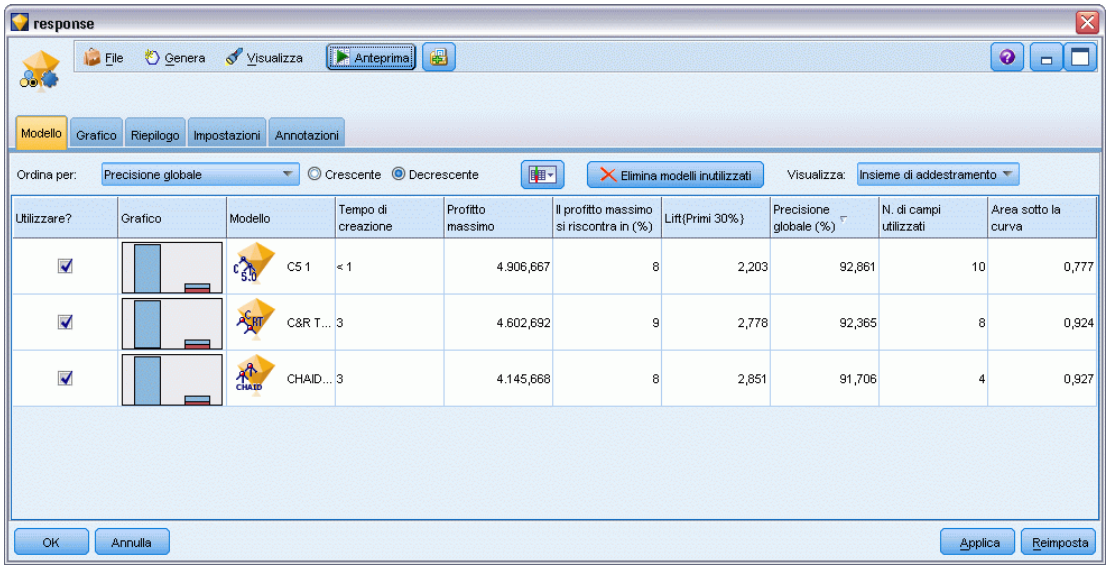
Dopo qualche minuto, l'insieme di modelli generato viene creato e collocato nell'area di disegno e nella palette Modelli nell'angolo superiore destro della finestra. L'insieme di modelli si può visualizzare, salvare o implementare in numerosi altri modi.

Aprire l'insieme di modelli; all'interno sono elencati i dettagli dei singoli modelli creati durante l'esecuzione. (In una situazione reale, in cui centinaia di modelli vengono creati su un insieme di dati di grandi dimensioni, questa operazione potrebbe richiedere parecchie ore). Vedere Figura 4-1 a pag. 45.

Se si desidera esplorare ulteriormente uno qualsiasi dei modelli, fare doppio clic sull'insieme di modelli nella colonna Modello per eseguire il drill-down e visualizzare i risultati del singolo modello. Da qui è possibile generare nodi Modelli, insiemi di modelli o grafici di valutazione.

Per generare un grafico a schermo intero è possibile fare doppio clic su un'anteprima nella colonna Grafico.

Figura 4-12
Risultati del nodo Classificatore automatico



Per default, i modelli vengono ordinati in base alla precisione globale, perché questa era la misurazione selezionata nella scheda Modelli del nodo Classificatore automatico. Il modello C51 genera una classificazione ottimale sulla base di questa misura, ma i modelli C&R Tree e CHAID sono ugualmente accurati.

È possibile effettuare l'ordinamento in base a una colonna diversa facendo clic sull'intestazione della colonna, oppure scegliere la misurazione desiderata dall'elenco a discesa Ordina per nella barra degli strumenti.

In base a questi risultati, è possibile decidere se utilizzare tutti e tre questi modelli accurati. Se si combinano le previsioni di più modelli è possibile superare le limitazioni dei singoli modelli e ottenere un livello di precisione complessiva più elevato.

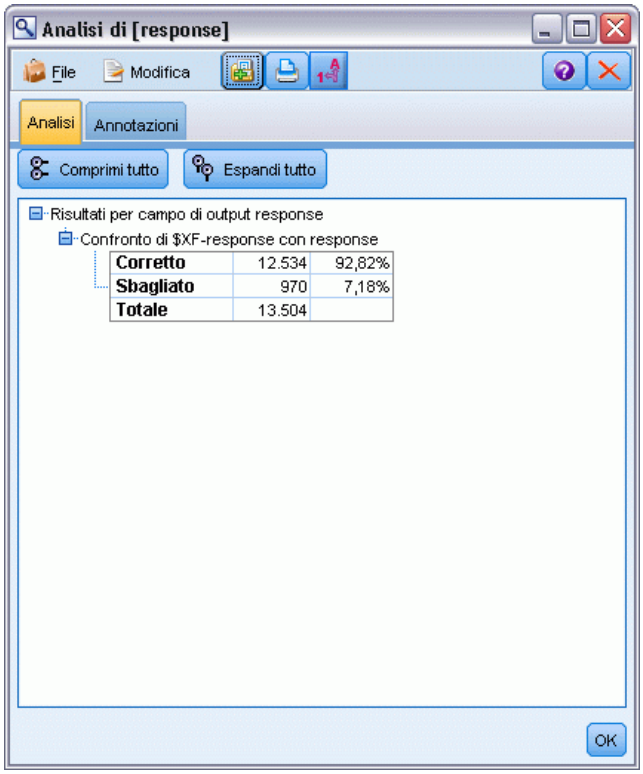
Nella colonna Uso?, selezionare i modelli C51, C&R Tree e CHAID.

Collegare un nodo Analisi (palette Output) dopo l'insieme di modelli. Fare clic con il pulsante destro del mouse sul nodo Analisi e scegliere Esegui per eseguire lo stream.

Il punteggio aggregato generato dai risultati classificatore binario viene aggiunto in un campo denominato *$XF-risposta*. Quando vengono misurati rispetto ai dati di addestramento, i valori previsti corrispondono alle risposte effettive (come registrate nel campo *risposta* originale) con una precisione globale pari a 92,82%.

Sebbene non sia altrettanto accurato del migliore dei tre modelli singoli di questo caso (92,86% per C51), la differenza è troppo piccola per essere significativa. In termini generali, un modello Risultati classificatore binario avrà maggiore probabilità di offrire buoni risultati se applicato a insiemi di dati diversi dai dati di addestramento.

Figura 4-13
Analisi dei tre modelli Classificatore binario



Riepilogo

Per riepilogare, è stato utilizzato il nodo Classificatore automatico per confrontare più modelli diversi, sono stati utilizzati i tre modelli più precisi che quindi sono stati aggiunti allo stream, in un insieme di modelli Classificatore automatico.

- In termini di precisione globale, i modelli C51, C&R Tree e CHAID hanno fornito le migliori prestazioni rispetto ai dati di addestramento.
- Il modello Classificatore binario ha fornito una prestazione simile a quella del migliore dei modelli singoli e potrebbe garantire prestazioni migliori se applicato ad altri insiemi di dati. Se l'obiettivo è quello di automatizzare il più possibile il processo, questo approccio consente di ottenere un modello robusto nella maggior parte delle situazioni, senza doversi addentrare nelle specifiche di ogni singolo modello.

Capitolo
5

Modellazione automatica per un obiettivo continuo

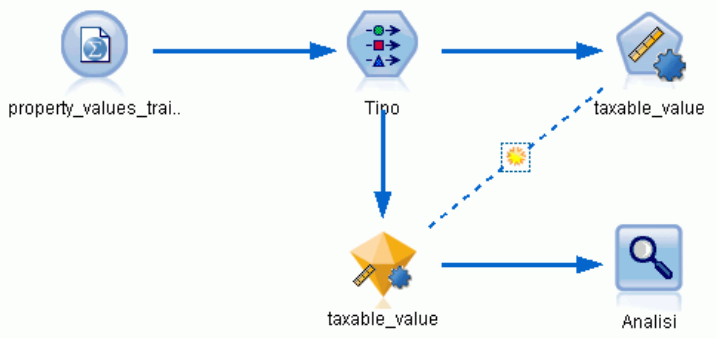
Valori di proprietà (Numerico automatico)

Il nodo Numerico automatico consente di creare e confrontare automaticamente diversi modelli per risultati continui (intervallo numerico), per esempio per prevedere il valore imponibile di una proprietà. Con un singolo nodo è possibile stimare e confrontare un insieme di modelli candidati e generare un sottoinsieme di modelli per analisi approfondite. Il nodo funziona analogamente al nodo Classificatore automatico, ma per obiettivi continui anziché per obiettivi flag o nominali.

Il nodo combina i modelli candidati migliori in un singolo modello aggregato utilizzando un insieme di modelli Risultati classificatore binario. Questo approccio unisce la facilità di automazione al vantaggio di combinare più modelli, che spesso generano previsioni più accurate di quelle che si otterrebbero con un solo modello.

Questo esempio è incentrato su un'amministrazione territoriale fittizia incaricata di rettificare e valutare le imposte sugli immobili. Per effettuare questa operazione con maggiore precisione, viene creato un modello che prevede i valori delle proprietà in base al tipo di edificio, al quartiere, alle dimensioni e ad altri fattori noti.

Figura 5-1
Stream di esempio di Numerico automatico



Questo esempio utilizza lo stream *property_values_numericpredictor.str*, installato nella cartella Demos in *streams*. Il file di dati utilizzato è *property_values_train.sav.* Per ulteriori informazioni, vedere l'argomento Cartella Demos in il capitolo 1 a pag. 7.

Dati di addestramento

Il file di dati include un campo denominato *valore_imponibile*, che è il **campo obiettivo**, o valore, che si desidera prevedere. Gli altri campi contengono informazioni quali quartiere, tipo di immobile e volume interno e possono essere utilizzati come predittori.

Nome del campo	Label
id_proprietà	ID della proprietà
quartiere	Area all'interno della città
tipo_edificio	Tipo di edificio
anno_costruzione	Anno di costruzione
volume_interno	Volume degli spazi interni
volume_altro	Volume degli annessi
dimensione_lotto	Dimensione del lotto
valore_imponibile	Valore imponibile

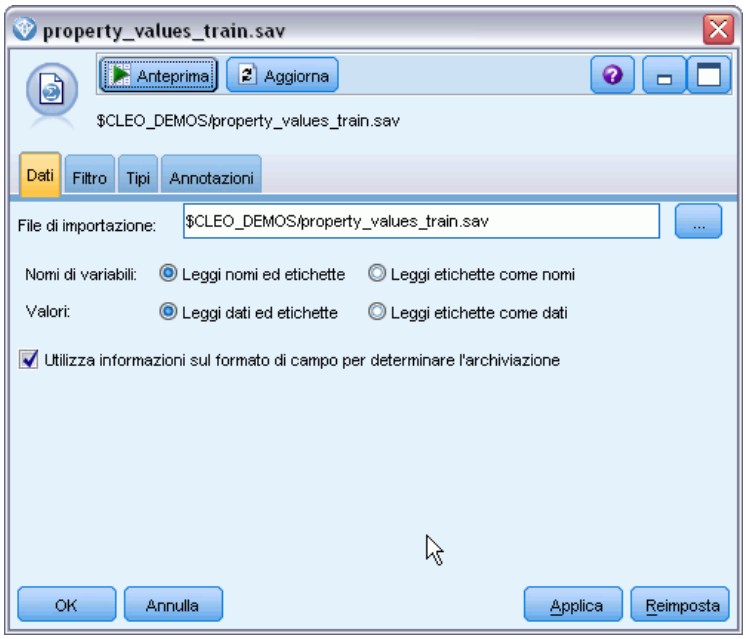
Nella cartella Demos è incluso inoltre un file dei dati di calcolo del punteggio denominato *property_values_score.sav*. Questo file contiene gli stessi campi, escluso il campo *valore_imponibile*. Dopo l'addestramento dei modelli con un insieme di dati in cui è noto il valore imponibile, è possibile calcolare il punteggio dei record nei quali questo valore non è ancora noto.

Creazione dello stream

- Aggiungere un nodo di input File Statistics che punti al file *property_values_train.sav* situato nella directory *Demos* dell'installazione IBM® SPSS® Modeler. (In alternativa, è possibile specificare

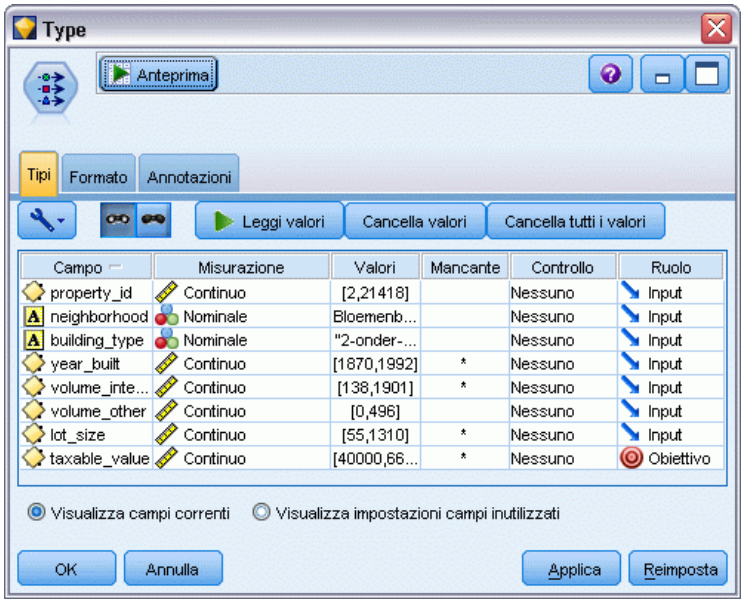
$CLEO_DEMOS/ nel percorso del file come collegamento per fare riferimento a questa cartella. Si noti che nel percorso è necessario utilizzare la barra e non la barra rovesciata, come mostrato.)

Figura 5-2
Lettura dei dati



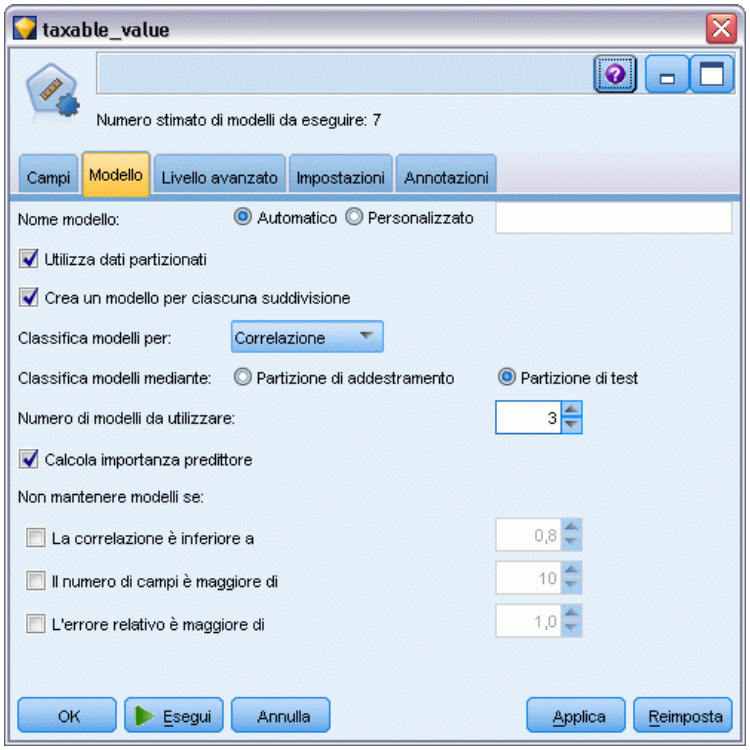
- Aggiungere un nodo Tipo e selezionare *valore_imponibile* come campo obiettivo (Ruolo = Obiettivo). Il ruolo degli altri campi deve essere impostato su Input in modo che vengano utilizzati come predittori.

Figura 5-3
Impostazione del campo obiettivo



- Collegare un nodo Numerico automatico e selezionare Correlazione come metrica per classificare i modelli.
- Impostare Numero di modelli da utilizzare su 3. Questo significa che quando si esegue il nodo vengono creati i tre modelli migliori.

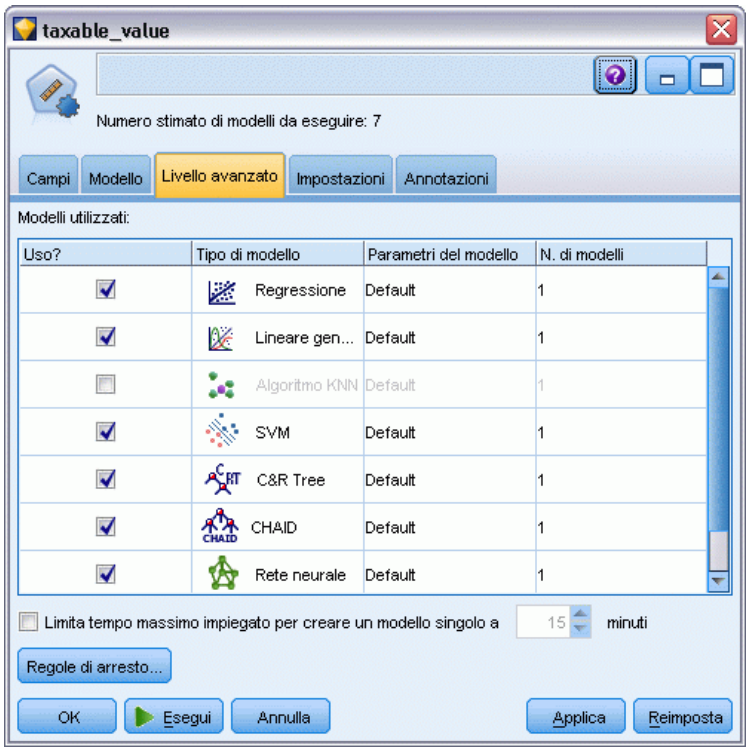
Figura 5-4
Scheda Modello del nodo Numerico automatico



- Nella scheda Livello avanzato, lasciare invariate le impostazioni di default; il nodo stima un singolo modello per ogni algoritmo, per un totale di sette modelli. (In alternativa, è possibile modificare queste impostazioni per confrontare più varianti per ogni tipo di modello).

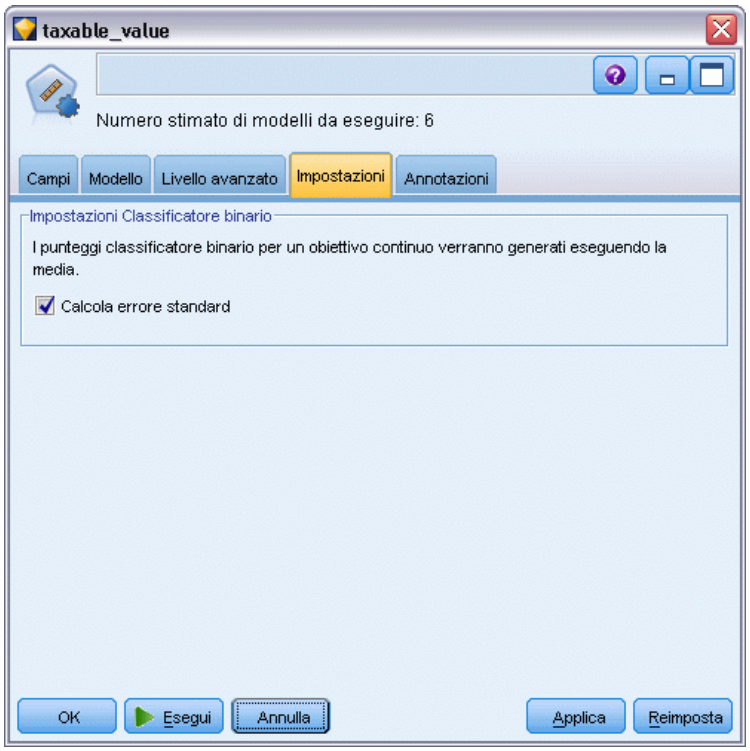
Poiché il Numero di modelli da utilizzare nella scheda Modelli è stato impostato su 3, il nodo calcolerà la precisione dei sette algoritmi e crea un unico insieme di modelli contenente i tre più accurati.

Figura 5-5
Scheda Livello avanzato del nodo Numerico automatico



▶ Nella scheda Impostazioni, lasciare le impostazioni di default. Trattandosi di un obiettivo continuo, il punteggio del nodo Risultati classificatore binario viene generato calcolando la media dei punteggi dei singoli modelli.

Figura 5-6
Scheda Impostazioni del nodo Numerico automatico



Confronto tra i modelli

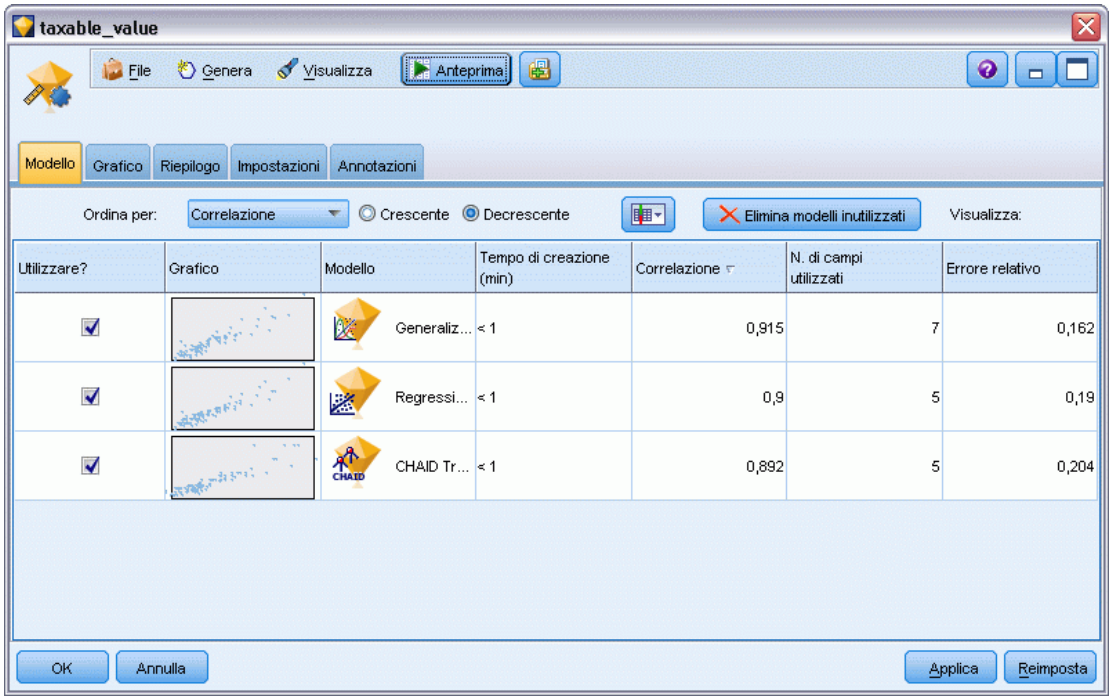
▶ Fare clic sul pulsante Esegui.

L'insieme di modelli viene creato e collocato nell'area di disegno, nonché nella palette Modelli nell'angolo superiore destro della finestra. L'insieme di modelli si può visualizzare, salvare o implementare in numerosi altri modi.

Aprire l'insieme di modelli; all'interno sono elencati i dettagli dei singoli modelli creati durante l'esecuzione. (In una situazione reale, in cui centinaia di modelli vengono stimati su un insieme di dati di grandi dimensioni, questa operazione potrebbe richiedere parecchie ore). Vedere Figura 5-1 a pag. 57.

Se si desidera esplorare ulteriormente uno qualsiasi dei modelli, fare doppio clic sull'insieme di modelli nella colonna Modello per eseguire il drill-down e visualizzare i risultati del singolo modello. Da qui è possibile generare nodi Modelli, insiemi di modelli o grafici di valutazione.

Figura 5-7
Risultati Numerico automatico



Per default, i modelli vengono ordinati in base alla correlazione poiché questa era la misura selezionata nel nodo Numerico automatico. Ai fini della classificazione si utilizza il valore assoluto della correlazione, in cui i valori più vicini a 1 indicano una relazione più forte. Il modello lineare generalizzato genera una classificazione ottimale sulla base di questa misura, ma molti altri modelli sono ugualmente accurati. Inoltre, il modello lineare generalizzato presenta l'errore relativo minimo.

È possibile effettuare l'ordinamento in base a una colonna diversa facendo clic sull'intestazione della colonna oppure scegliere la misura desiderata dall'elenco Ordina per sulla barra degli strumenti.

Ogni grafico rappresenta graficamente i valori osservati rispetto a quelli previsti per il modello e fornisce una rapida indicazione visiva della loro correlazione. In un modello efficace, i punti si raggruppano lungo la diagonale, cosa che avviene per tutti i modelli di questo esempio.

Per generare un grafico a schermo intero è possibile fare doppio clic su un'anteprima nella colonna Grafico.

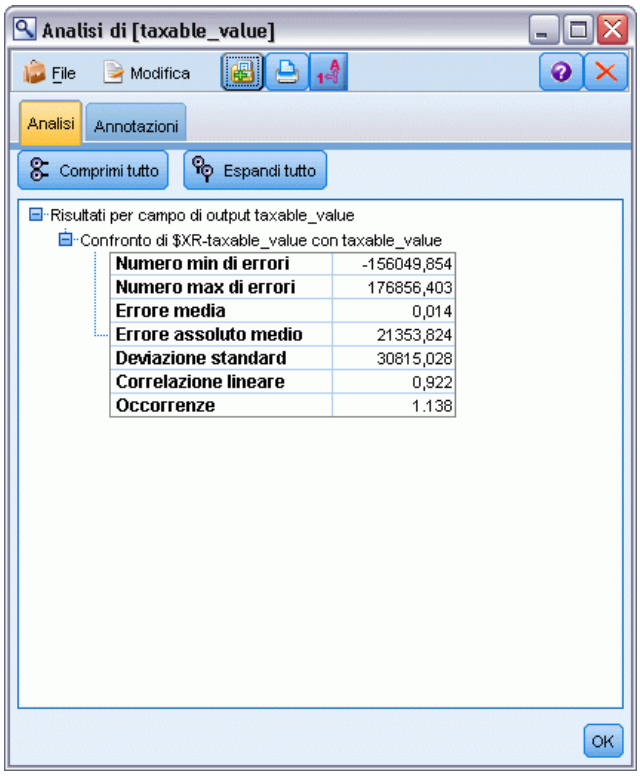
In base a questi risultati, è possibile decidere se utilizzare tutti e tre questi modelli accurati. Se si combinano le previsioni di più modelli è possibile superare le limitazioni dei singoli modelli e ottenere un livello di precisione complessiva più elevato.

Nella colonna Utilizzare?, verificare che tutti e tre i modelli siano selezionati.

Collegare un nodo Analisi (palette Output) dopo l'insieme di modelli. Fare clic con il pulsante destro del mouse sul nodo Analisi e scegliere Esegui per eseguire lo stream.

Il punteggio medio generato dal nodo Risultati classificatore binario viene aggiunto in un campo denominato *$XR-valore_imponibile*, con una correlazione di 0,922, che è superiore a quelle dei tre modelli singoli. I punteggi del nodo Risultati classificatore binario indicano un basso errore assoluto medio e possono generare prestazioni migliori rispetto ai singoli modelli se applicati ad altri insiemi di dati.

Figura 5-8
Stream di esempio di Numerico automatico



Riepilogo

Per riepilogare, è stato utilizzato il nodo Numerico automatico per confrontare più modelli diversi, sono stati selezionati i tre modelli più precisi che quindi sono stati aggiunti allo stream, in un insieme di modelli Numerico automatico aggregato.

- In termini di precisione globale, i modelli lineari generalizzati, di regressione e CHAID hanno fornito le migliori prestazioni rispetto ai dati di addestramento.
- Il modello Risultati classificatore binario ha mostrato una prestazione superiore a quella di due dei singoli modelli e potrebbe generare prestazioni migliori se applicato ad altri insiemi di dati. Se l'obiettivo è quello di automatizzare il più possibile il processo, questo approccio consente di ottenere un modello robusto nella maggior parte delle situazioni, senza doversi addentrare nelle specifiche di ogni singolo modello.

Parte II:
Esempi di preparazione di dati

Capitolo

6

Preparazione automatica dei dati (ADP)

La preparazione dei dati per l'analisi è uno degli aspetti più importanti in qualsiasi progetto di data mining e in genere richiede tempi molto lunghi. Il nodo Preparazione automatica dati (ADP) gestisce automaticamente questa operazione, analizzando i dati e individuando le correzioni, escludendo i campi problematici o probabilmente inutili e derivando all'occorrenza nuovi attributi, migliorando inoltre le performance grazie alle tecniche di screening intelligente. Il nodo può essere utilizzato in modo completamente automatico, per esempio lasciando al nodo la scelta e l'applicazione delle correzioni, oppure è possibile visualizzare in anteprima le modifiche prima di apportarle accettandole o respingendole a seconda dei casi.

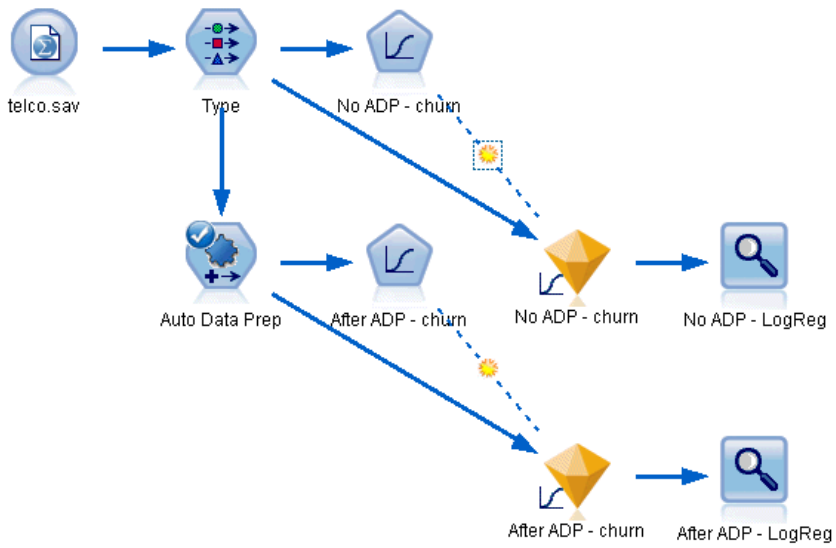
L'uso del nodo ADP consente di preparare rapidamente e con facilità i dati per il data mining senza che siano necessarie conoscenze pregresse sui concetti statistici coinvolti. Se si esegue il nodo con le impostazioni di default, la creazione e il calcolo del punteggio dei modelli tenderanno a essere più rapidi.

Questo esempio utilizza lo stream *ADP_basic_demo.str* che fa riferimento al file di dati *telco.sav* per dimostrare la maggiore precisione ottenibile mantenendo le impostazioni di default del nodo ADP per la creazione dei modelli. Questi file sono disponibili nella directory *Demos* di tutte le installazioni di IBM® SPSS® Modeler. L'accesso si effettua dal gruppo di programmi IBM® SPSS® Modeler nel menu Start di Windows. Il file *ADP_basic_demo.str* si trova nella directory *streams*.

Creazione dello stream

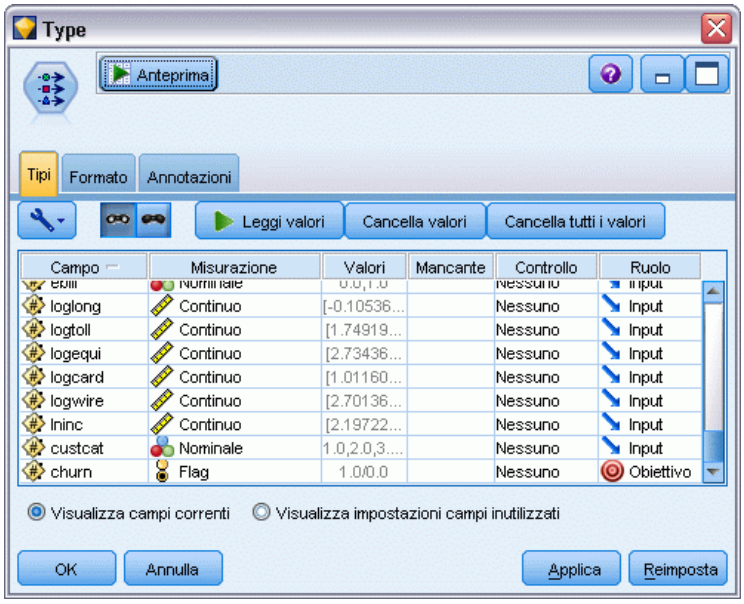
- Per creare lo stream, aggiungere un nodo File Statistics di input che punti al file *telco.sav* situato nella directory *Demos* dell'installazione IBM® SPSS® Modeler.

Figura 6-1
Creazione dello stream



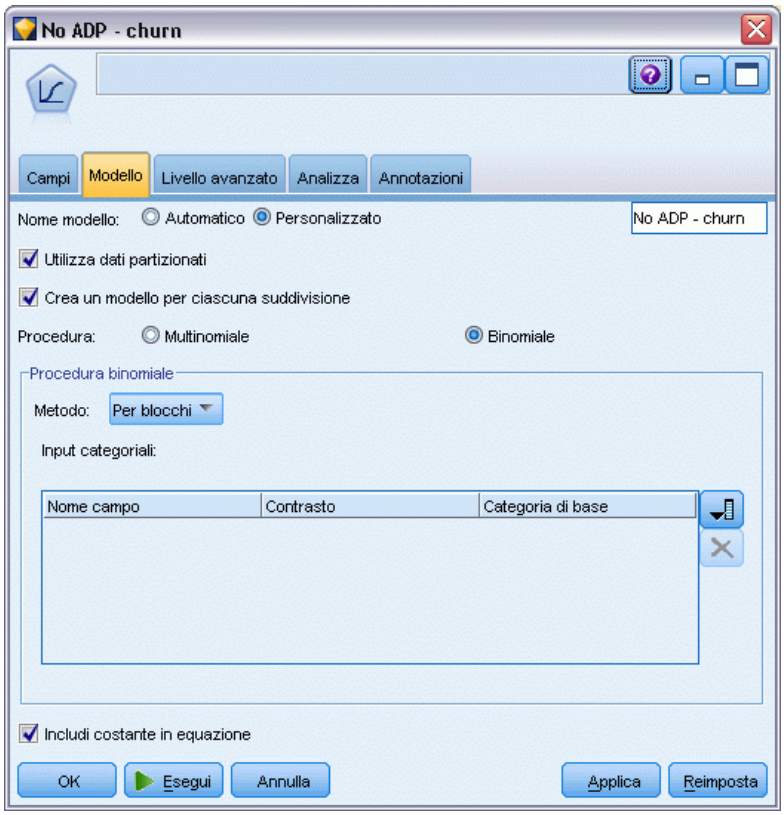
- Collegare un nodo Tipo al nodo di input, impostare il livello di misurazione per il campo *Tasso di abbandono* su Flag e impostare il ruolo su Obiettivo. Tutti gli altri campi dovrebbero avere il ruolo impostato su Input.

Figura 6-2
Selezione dell'obiettivo



- Collegare un nodo Logistica al nodo Tipo.
- Nel nodo Logistica, fare clic sulla scheda Modello e selezionare la procedura Binomiale. Nel campo *Nome modello*, selezionare Personalizzato e immettere No ADP - tasso di abbandono.

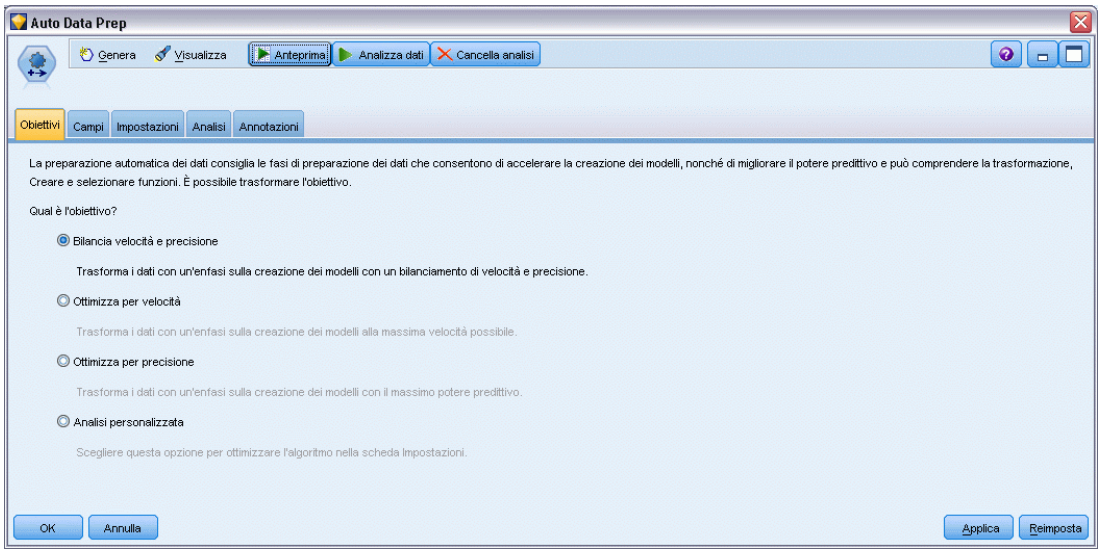
Figura 6-3
Scelta delle opzioni del modello



- Collegare un nodo ADP al nodo Tipo. Nella scheda Obiettivi, mantenere le impostazioni di default per analizzare e preparare i dati con un rapporto equilibrato tra velocità e precisione.
- Fare clic su Analizza dati nella parte superiore della scheda Obiettivi per analizzare ed elaborare i dati.

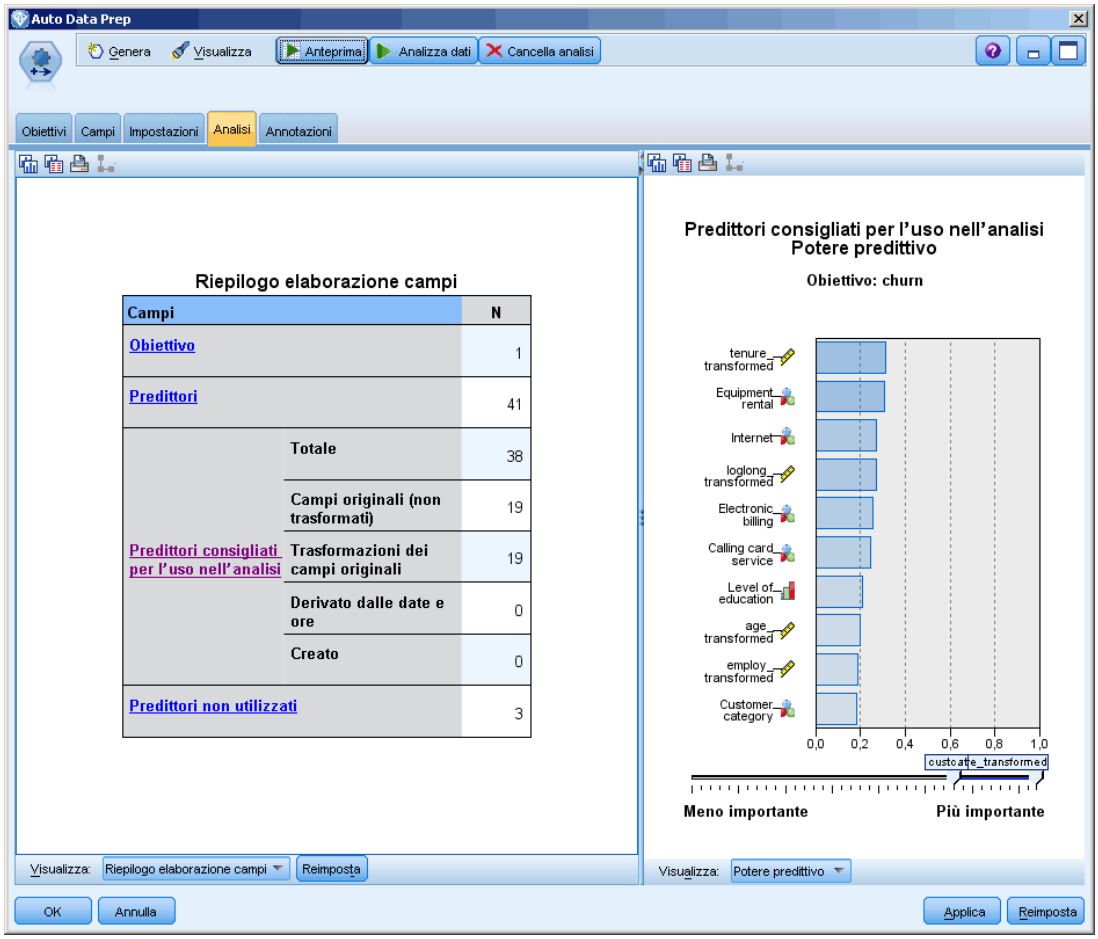
Le altre opzioni per il nodo ADP consentono di indicare se si desidera concentrarsi maggiormente sulla precisione o sulla velocità di elaborazione, oppure di ottimizzare molte fasi di elaborazione della preparazione dei dati.

Figura 6-4
Obiettivi di default del nodo ADP



I risultati dell'elaborazione dei dati sono visualizzati nella scheda Analisi. Il Riepilogo elaborazione campi mostra che, su 41 funzioni dei dati inserite nel nodo ADP, 19 sono state trasformate per facilitare l'elaborazione e 3 sono state scartate perché non utilizzate.

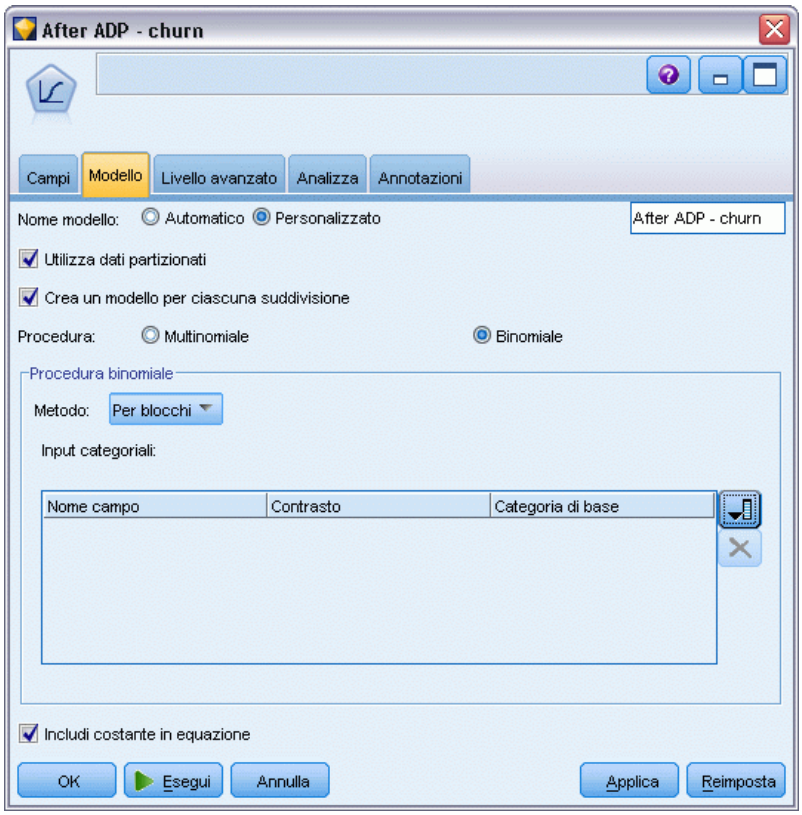
Figura 6-5
Riepilogo dell'elaborazione dati



► Collegare un nodo Logistica al nodo ADP.

▶ Nel nodo Logistica, fare clic sulla scheda Modello e selezionare la procedura Binomiale. Nel campo *Nome modello*, selezionare Personalizzato e immettere Dopo ADP - tasso di abbandono.

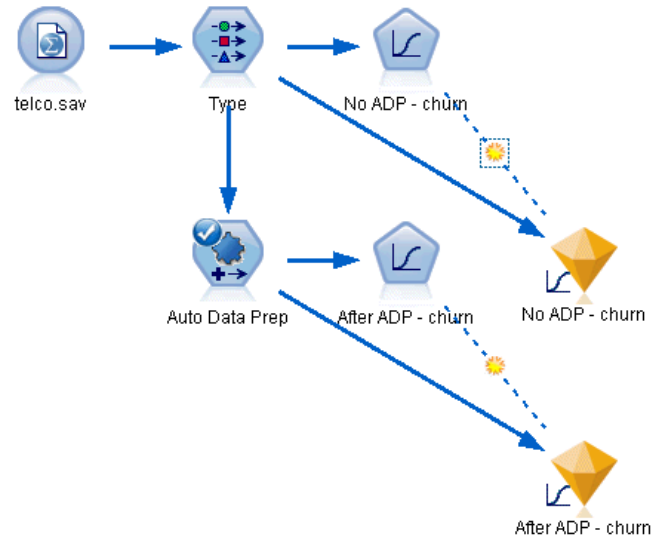
Figura 6-6
Scelta delle opzioni del modello



Confronto della precisione dei modelli

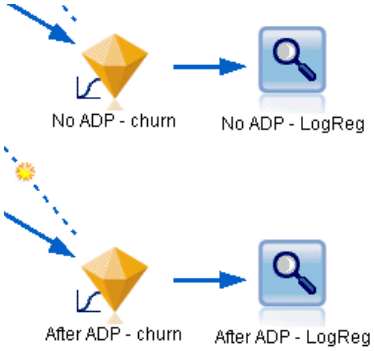
- Eseguire entrambi i nodi Logistica per generare gli insiemi di modelli, che saranno aggiunti allo stream e alla palette Modelli nell'angolo superiore destro.

Figura 6-7
Collegamento degli insiemi di modelli



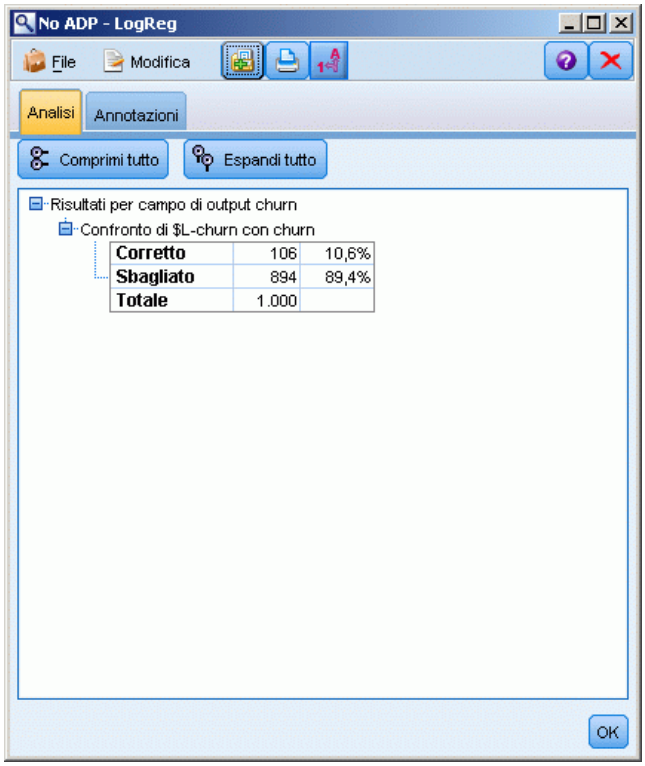
- Collegare dei nodi Analisi agli insiemi di modelli ed eseguire i nodi Analisi utilizzando le rispettive impostazioni di default.

Figura 6-8
Collegamento dei nodi Analisi



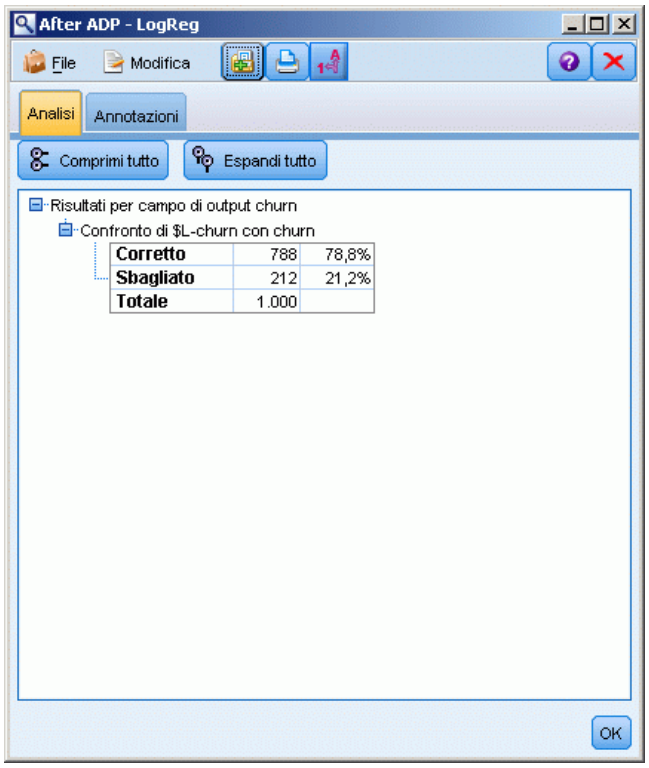
L'analisi del modello non ottenuto mediante il nodo ADP mostra che la semplice elaborazione dei dati mediante il nodo Regressione logistica con le relative impostazioni di default genera un modello di scarsa precisione: solo il 10,6%.

Figura 6-9
Risultati del modello non generato mediante il nodo ADP



L'analisi del modello ottenuto mediante il nodo ADP mostra che l'elaborazione dei dati con le impostazioni ADP di default genera un modello molto più preciso, corretto per il 78,8%.

Figura 6-10
Risultati del modello generato mediante il nodo ADP



Per riassumere, con la semplice esecuzione del nodo ADP per ottimizzare l'elaborazione dei dati è stato possibile creare un modello più preciso con una manipolazione diretta dei dati limitata.

Ovviamente, se quello che interessa è dimostrare la validità o meno di una determinata teoria o creare dei modelli specifici, può essere utile lavorare direttamente con le impostazioni del modello; se invece il tempo a disposizione è poco o se si dispone di una grande quantità di dati da preparare, il nodo ADP può rivelarsi vantaggioso.

Per una spiegazione dei fondamenti matematici dei metodi di modellazione utilizzati in IBM® SPSS® Modeler, vedere il manuale *SPSS Modeler Algorithms Guide*, disponibile nella directory *\Documentation* del disco di installazione.

Si noti che i risultati di questo esempio si basano soltanto sui dati di addestramento. Per un'indicazione dell'efficacia con cui i modelli potranno essere estesi ad altri dati in una situazione reale, utilizzare un nodo Partizione per estrarre un sottoinsieme di record da utilizzare per test e validazione. Per ulteriori informazioni, vedere l'argomento Nodo Partizione in il capitolo 4 in *IBM SPSS Modeler 15 Nodi di input, elaborazione e output*.

Capitolo

7

Preparazione dei dati per l'analisi (Esplorazione dei dati)

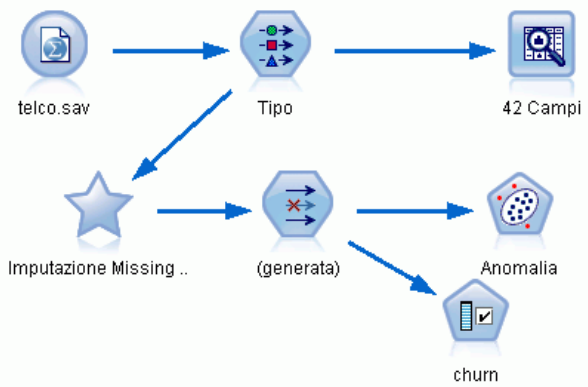
Il nodo Esplora offre una prima panoramica completa dei dati inseriti in IBM® SPSS® Modeler. Utilizzato frequentemente durante l'esplorazione iniziale dei dati, il report del nodo Esplora contiene statistiche riassuntive, nonché istogrammi e grafici di distribuzione per ogni campo di dati e consente di specificare il trattamento dei valori mancanti, anomali ed estremi.

In questo esempio viene utilizzato lo stream denominato *telco_dataaudit.str*, che fa riferimento al file di dati denominato *telco.sav*. Questi file sono disponibili nella directory *Demos* di tutte le installazioni di IBM® SPSS® Modeler. L'accesso si effettua dal gruppo di programmi SPSS Modeler nel menu Start di Windows. Il file *telco_dataaudit.str* si trova nella directory *streams*.

Creazione dello stream

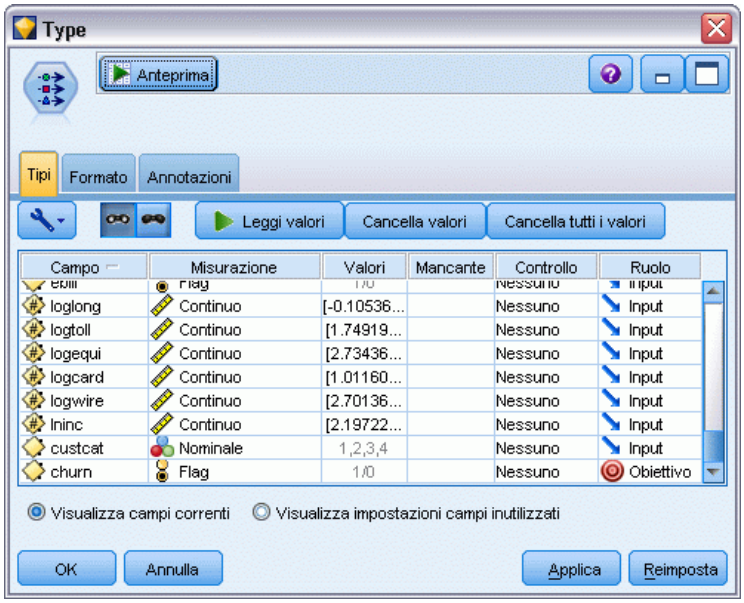
- Per creare lo stream, aggiungere un nodo File Statistics di input che punti al file *telco.sav* situato nella directory *Demos* dell'installazione IBM® SPSS® Modeler.

Figura 7-1
Creazione dello stream



- Aggiungere un nodo Tipo per definire i campi e specificare *tasso di abbandono* come campo obiettivo (Ruolo = Obiettivo). Il ruolo dovrebbe essere impostato su Input per tutti gli altri campi in modo che questo rappresenti l'unico obiettivo.

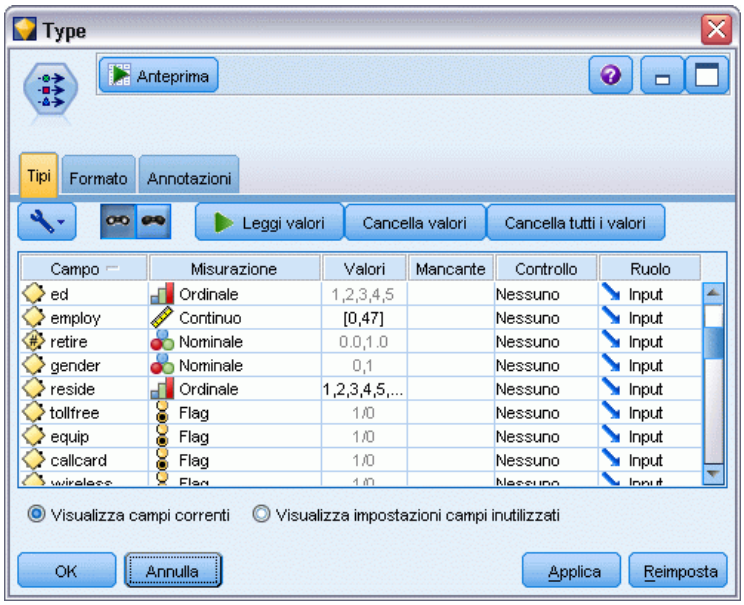
Figura 7-2
Impostazione dell'obiettivo



- Verificare che i livelli di misurazione dei campi siano definiti correttamente. Per esempio, la maggior parte dei campi con valore 0 e 1 può essere considerata di tipo flag, ma alcuni campi,

come quello relativo al sesso, vengono visualizzati più accuratamente come campi nominali con due valori.

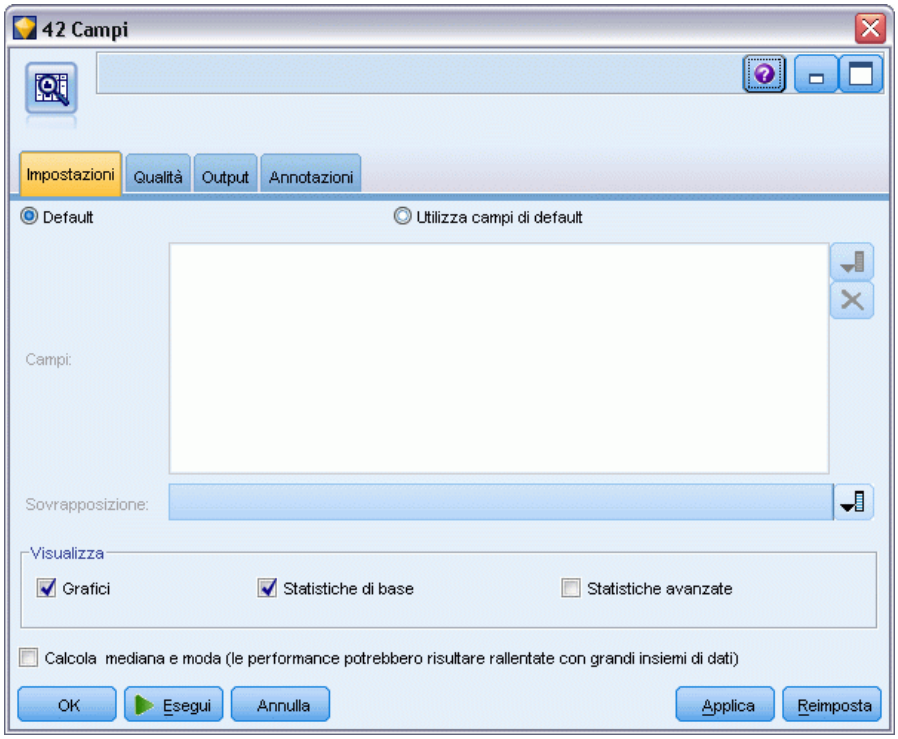
Figura 7-3
Impostazione dei livelli di misurazione



Suggerimento: se si desidera modificare le proprietà di più campi con valori simili (quali 0/1), fare clic sull'intestazione della colonna *Valori* per ordinare i campi in base a quella colonna e utilizzare il tasto Maiusc per selezionare tutti i campi da modificare. A questo punto, è possibile fare clic con il pulsante destro del mouse sulla selezione per cambiare il livello di misurazione o altri attributi di tutti i campi selezionati.

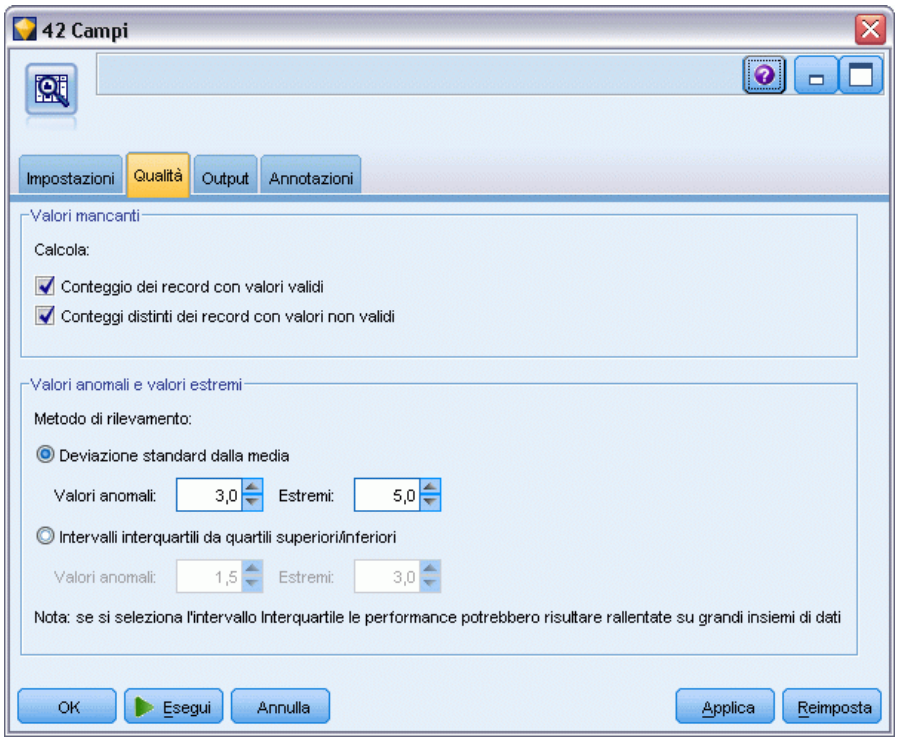
- Collegare un nodo Esplora allo stream. Nella scheda Impostazioni, mantenere le impostazioni di default in modo da includere tutti i campi nel report. Poiché *tasso di abbandono* è l'unico campo obiettivo definito nel nodo Tipo, verrà automaticamente utilizzato come sovrapposizione.

Figura 7-4
Scheda Impostazioni del nodo Esplora



Nella scheda Qualità, mantenere le impostazioni di default per il rilevamento dei valori mancanti, anomali ed estremi, quindi fare clic su Esegui.

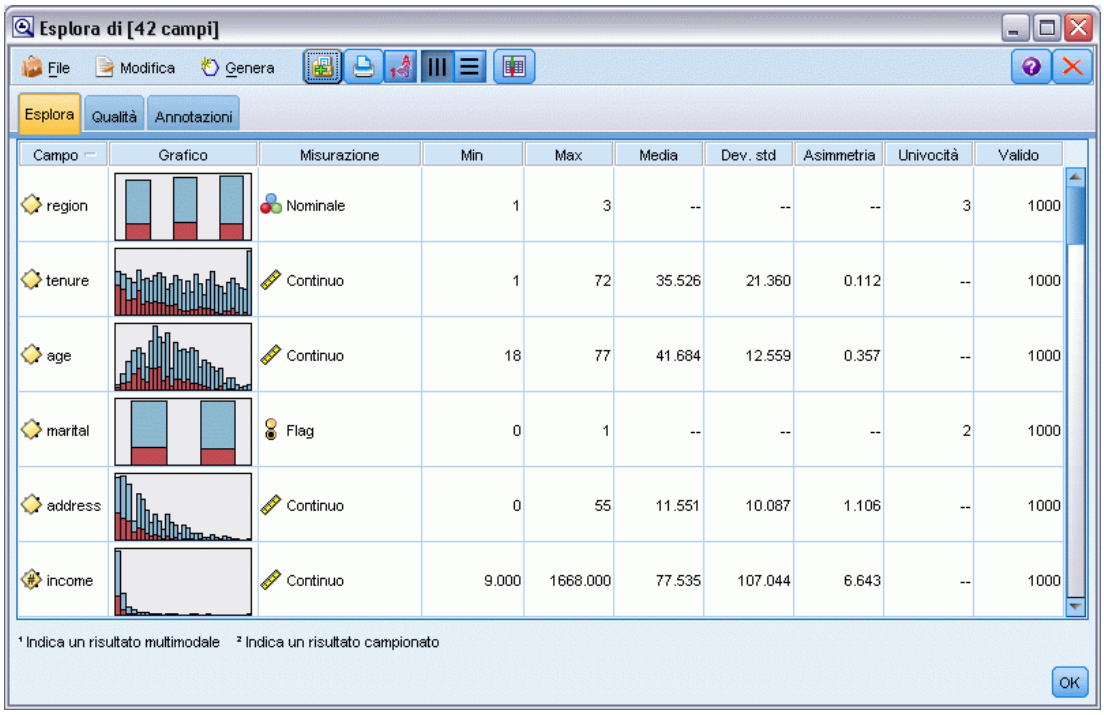
Figura 7-5
Scheda Qualità del nodo Esplora



Ricerca fra statistiche e diagrammi

Viene visualizzato il browser del nodo Esplora, con un'anteprima dei grafici e delle statistiche descrittive per ciascun campo.

Figura 7-6
Browser del nodo Esplora

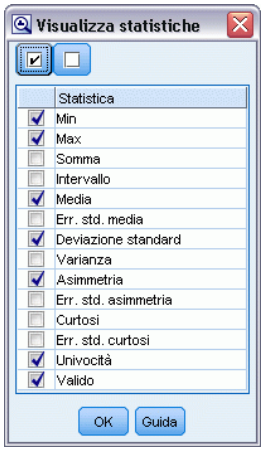


Campo	Grafico	Misurazione	Min	Max	Media	Dev. std	Asimmetria	Univocità	Valido
region		Nominale	1	3	--	--	--	3	1000
tenure		Continuo	1	72	35.526	21.360	0.112	--	1000
age		Continuo	18	77	41.684	12.559	0.357	--	1000
marital		Flag	0	1	--	--	--	2	1000
address		Continuo	0	55	11.551	10.087	1.106	--	1000
income		Continuo	9.000	1668.000	77.535	107.044	6.643	--	1000

Utilizzare la barra degli strumenti per visualizzare le etichette di campi e valori e per cambiare da orizzontale a verticale (e viceversa) l'allineamento dei diagrammi (solo per i campi categoriali).

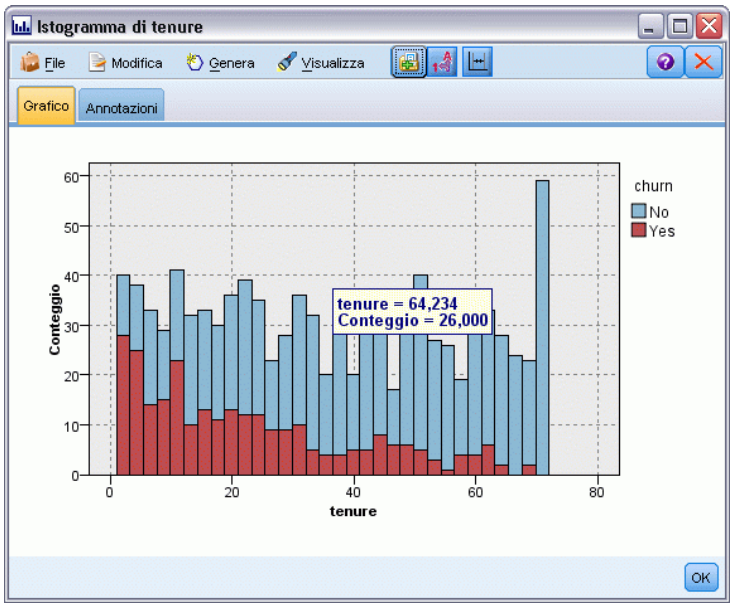
- È inoltre possibile utilizzare la barra degli strumenti o il menu Modifica per scegliere le statistiche che si desidera visualizzare.

Figura 7-7
Visualizza statistiche



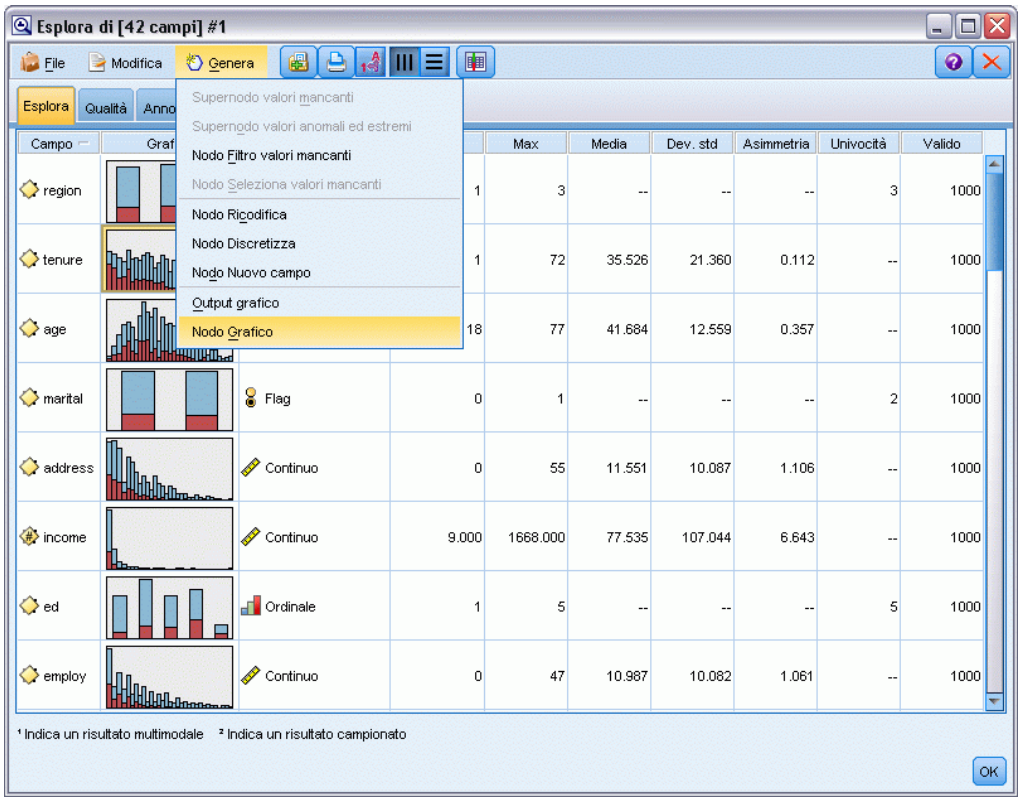
Fare doppio clic su un'anteprima di diagramma nel report di esplorazione per visualizzare lo stesso diagramma in dimensioni normali. Poiché *tasso di abbandono* è l'unico campo obiettivo presente nello stream, viene automaticamente utilizzato come sovrapposizione. Visualizzare e nascondere le etichette dei campi e dei valori mediante la barra degli strumenti della finestra del diagramma oppure fare clic sul pulsante Modifica per personalizzare ulteriormente il diagramma.

Figura 7-8
Istogramma di durata



In alternativa, è possibile selezionare una o più anteprime e generare un nodo Grafico per ognuna. I nodi generati vengono posti nell'area di disegno dello stream e possono essere aggiunti allo stream per ricreare quel grafico specifico.

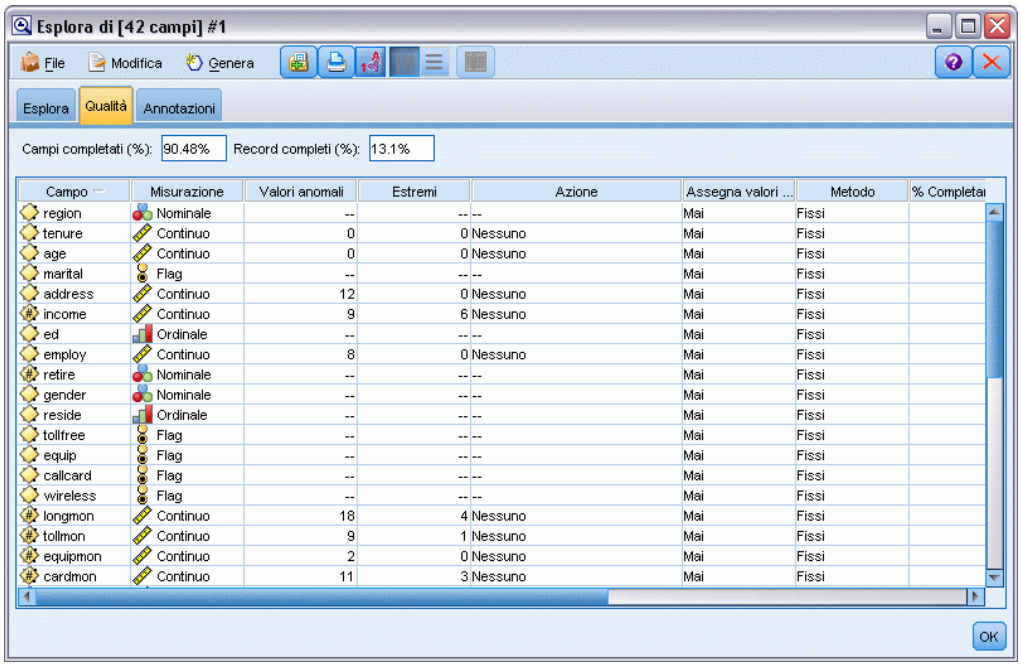
Figura 7-9
Generazione di un nodo Grafico



Gestione dei valori anomali e mancanti

La scheda Qualità nel report di esplorazione visualizza informazioni sui valori anomali, estremi e mancanti.

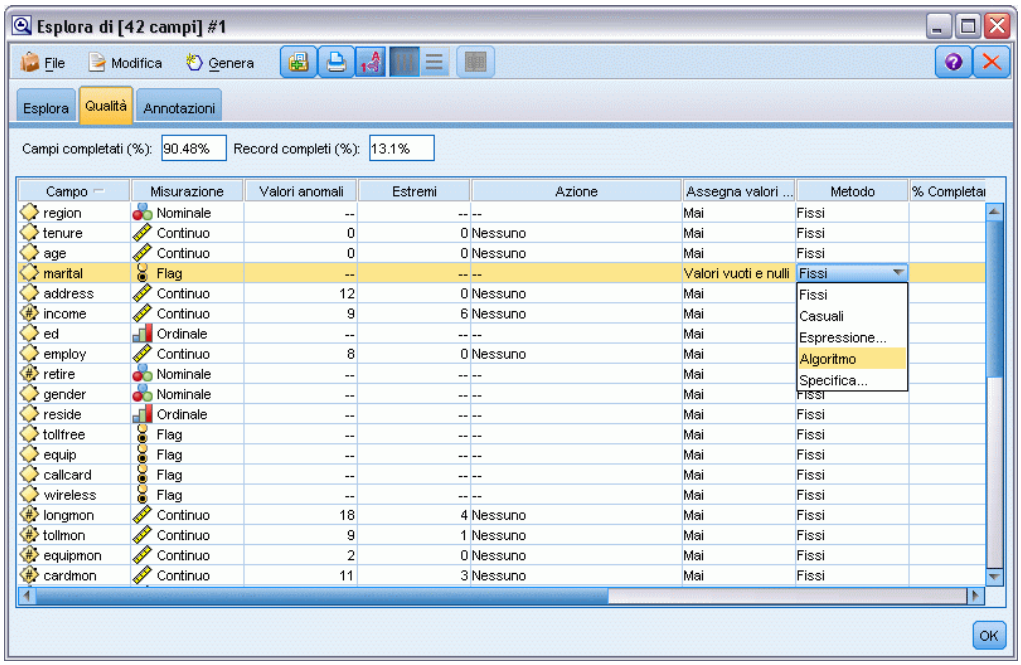
Figura 7-10
Scheda Qualità del browser Esplora



Campo	Misurazione	Valori anomali	Estremi	Azione	Assegna valori ...	Metodo	% Completai
region	Nominale	--	--	--	Mai	Fissi	
tenure	Continuo	0	0	Nessuno	Mai	Fissi	
age	Continuo	0	0	Nessuno	Mai	Fissi	
marital	Flag	--	--	--	Mai	Fissi	
address	Continuo	12	0	Nessuno	Mai	Fissi	
income	Continuo	9	6	Nessuno	Mai	Fissi	
ed	Ordinale	--	--	--	Mai	Fissi	
employ	Continuo	8	0	Nessuno	Mai	Fissi	
retire	Nominale	--	--	--	Mai	Fissi	
gender	Nominale	--	--	--	Mai	Fissi	
reside	Ordinale	--	--	--	Mai	Fissi	
tollfree	Flag	--	--	--	Mai	Fissi	
equip	Flag	--	--	--	Mai	Fissi	
callcard	Flag	--	--	--	Mai	Fissi	
wireless	Flag	--	--	--	Mai	Fissi	
longmon	Continuo	18	4	Nessuno	Mai	Fissi	
tollmon	Continuo	9	1	Nessuno	Mai	Fissi	
equipmon	Continuo	2	0	Nessuno	Mai	Fissi	
cardmon	Continuo	11	3	Nessuno	Mai	Fissi	

È anche possibile specificare i metodi di gestione di questi valori e creare Supernodi che applichino automaticamente le trasformazioni. Per esempio, è possibile selezionare uno o più campi e scegliere di assegnare o sostituire i valori mancanti per tali campi utilizzando diversi metodi, fra cui l'algoritmo C&RT.

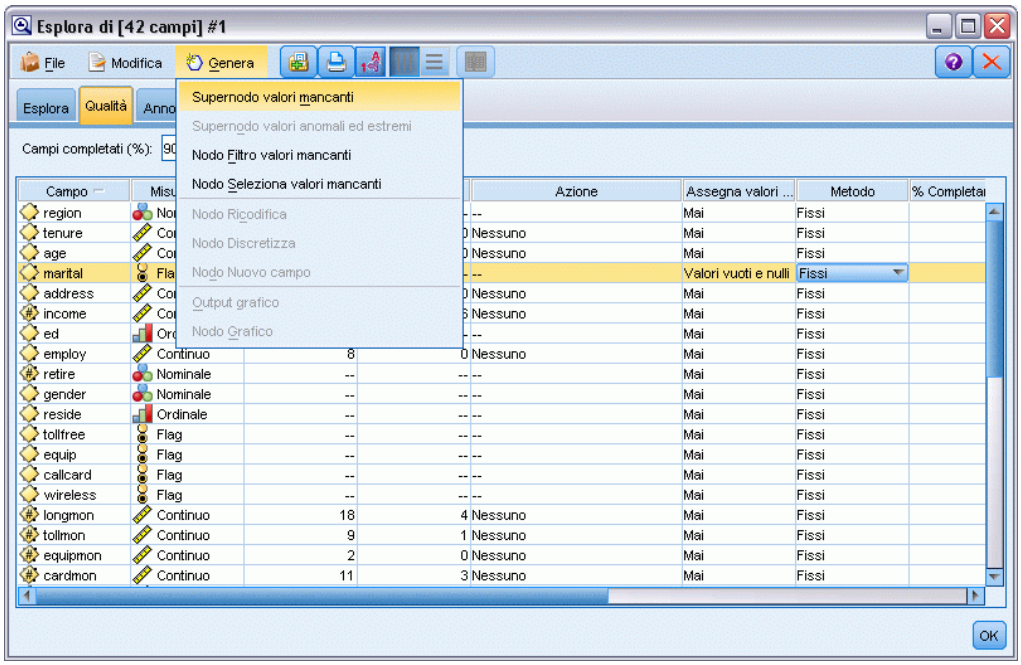
Figura 7-11
Scelta di un metodo di assegnazione



Dopo aver specificato il metodo di assegnazione per uno o più campi, per generare un Supernodo valori mancanti, dai menu scegliere:

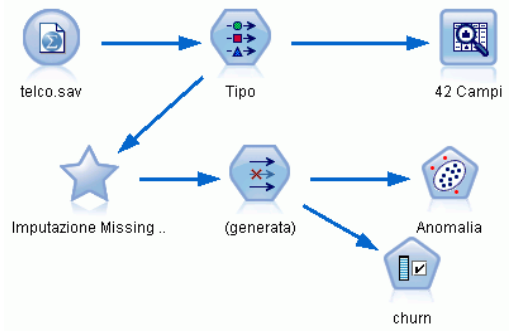
Genera > Supernodo valori mancanti

Figura 7-12
Generazione del Supernodo



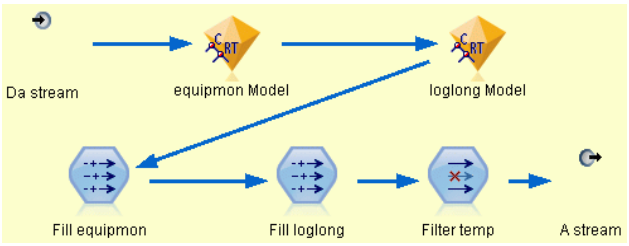
Il Supernodo generato viene aggiunto all'area di disegno dello stream, da dove sarà possibile collegarlo allo stream per applicare le trasformazioni.

Figura 7-13
Stream con Supernodo valori mancanti



Il Supernodo contiene una serie di nodi che eseguono le trasformazioni richieste. Per comprenderne il funzionamento, è possibile modificare il Supernodo e fare clic su Zoom avanti.

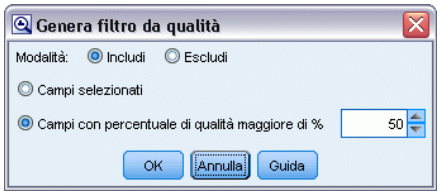
Figura 7-14
Zoom avanti nel Supernodo



Per ogni campo assegnato utilizzando il metodo dell'algoritmo, per esempio, ci sarà un modello C&RT diverso e un nodo Riempimento che sostituisce i valori vuoti e nulli con il valore previsto dal modello. È possibile aggiungere, modificare o rimuovere nodi specifici all'interno del Supernodo per personalizzare ulteriormente il comportamento.

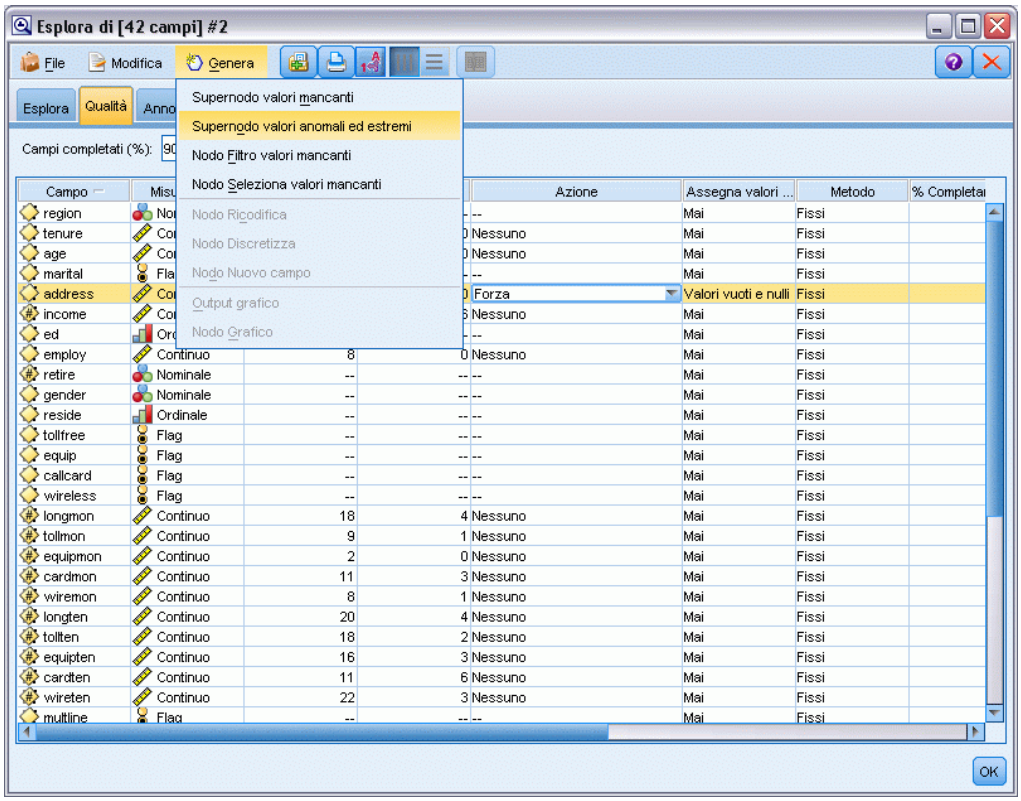
In alternativa, è possibile generare un nodo Seleziona o Filtro per rimuovere i campi o i record con valori mancanti. Per esempio, è possibile filtrare tutti i campi che presentano una percentuale di qualità inferiore a una soglia specifica.

Figura 7-15
Generazione di un nodo Filtro



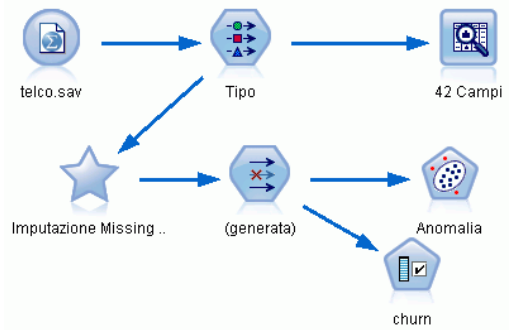
I valori anomali e i valori estremi possono essere gestiti in modo analogo. Specificare l'azione che si desidera eseguire per ciascuno dei campi (forzare, scartare o annullare), quindi generare un Supernodo per applicare le trasformazioni.

Figura 7-16
Generazione di un nodo Filtro



Al termine dell'esplorazione e dopo avere aggiunto allo stream i nodi generati, è possibile procedere con l'analisi. Se lo si desidera, è anche possibile sottoporre i dati a ulteriore screening mediante Rilevamento anomalie, Selezione funzioni e una serie di altri metodi.

Figura 7-17
Stream con Supernodo valori mancanti



Capitolo
8

Trattamenti farmacologici (Grafici preliminari/C5.0)

Si supponga di essere un ricercatore medico alle prese con la compilazione di dati per uno studio. Sono stati raccolti dati relativi a un gruppo di pazienti, tutti colpiti dalla stessa malattia. Nel corso della terapia, ogni paziente è stato sottoposto a una cura scelta tra cinque. Si desidera utilizzare quindi il data mining per individuare la cura più appropriata per un paziente che soffra della stessa malattia.

In questo esempio viene utilizzato lo stream denominato *druglearn.str*, che fa riferimento al file di dati denominato *DRUG1n*. Questi file sono disponibili nella directory *Demos* di tutte le installazioni di IBM® SPSS® Modeler. L'accesso si effettua dal gruppo di programmi IBM® SPSS® Modeler nel menu Start di Windows. Il file *druglearn.str* si trova nella directory *streams*.

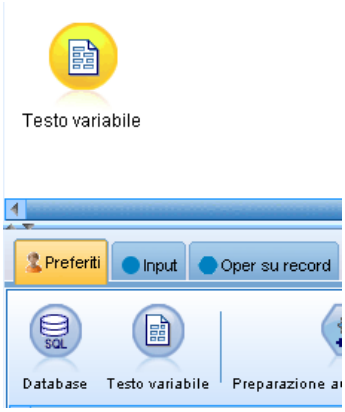
I campi di dati utilizzati per questo esempio sono i seguenti:

Campo dati	Descrizione
Età	(Numero)
Sesso	*M* o *F*
Pressione	Valori della pressione sanguigna: *ALTO*, *NORMALE* o *BASSO*
Colesterolo	Valori del colesterolo nel sangue: *NORMALE* o *ALTO*
Na	Concentrazione di sodio nel sangue
K	Concentrazione di potassio nel sangue
Cura	Cura prescritta a cui il paziente è stato sottoposto

Lettura di dati di testo

È possibile leggere i dati di testo utilizzando un **nodo Testo variabile**. È possibile aggiungere un nodo Testo variabile dalle palette selezionando la scheda Input per cercare il nodo oppure utilizzando la scheda Preferiti che include il nodo per default. Fare quindi doppio clic sul nuovo nodo per visualizzare la relativa finestra di dialogo.

Figura 8-1
Aggiunta di un nodo Testo variabile



Fare clic sul pulsante a destra della casella File contrassegnato dai puntini di sospensione (...) per passare alla directory nella quale è installato IBM® SPSS® Modeler nel sistema. Aprire la directory *Demos* e selezionare il file denominato *DRUG1n*.

Verificare che sia selezionato Leggi i nomi dei campi dal file e controllare i campi e i valori appena caricati nella finestra di dialogo.

Figura 8-2
Finestra di dialogo Testo variabile

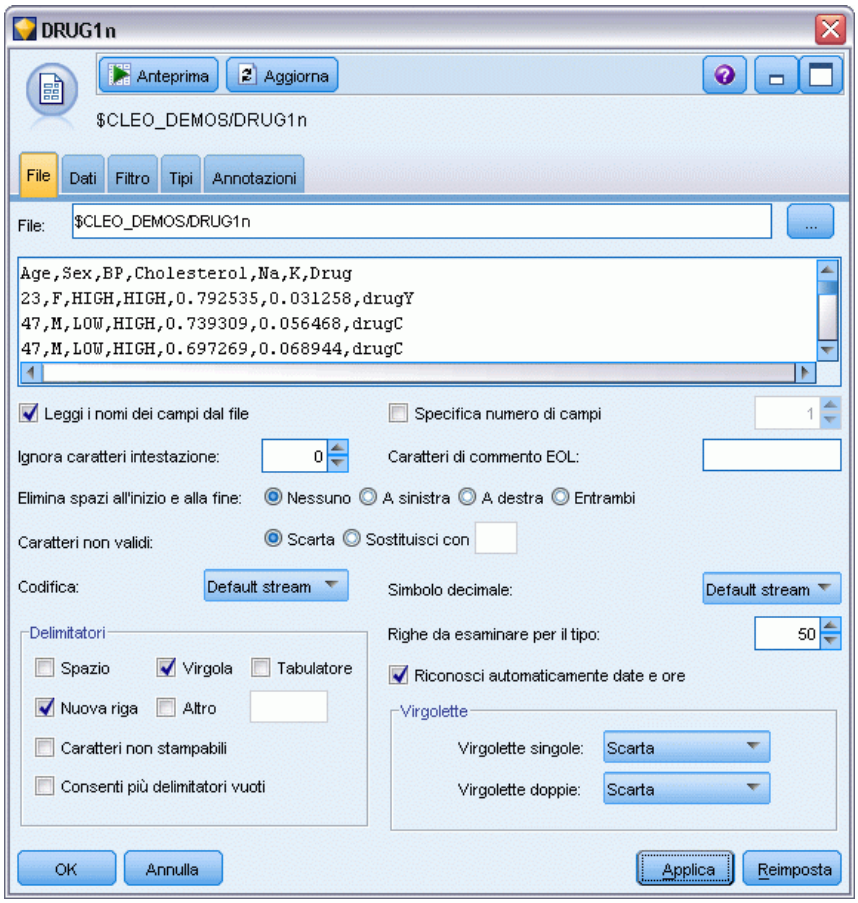


Figura 8-3
Modifica del tipo di archiviazione per un campo

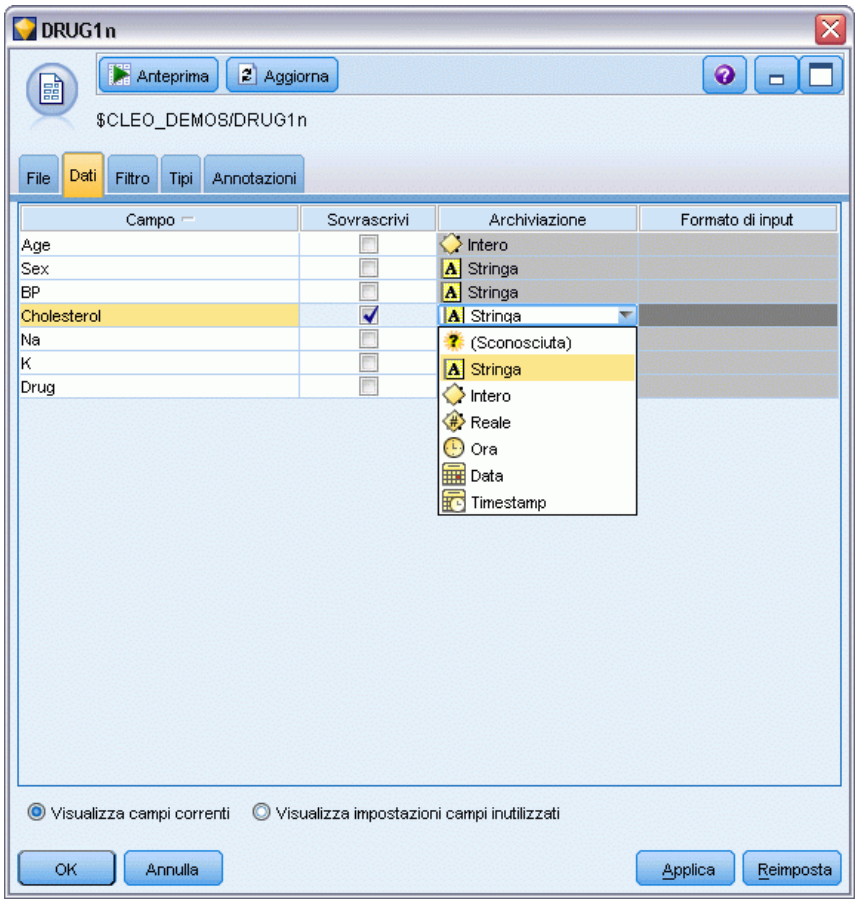
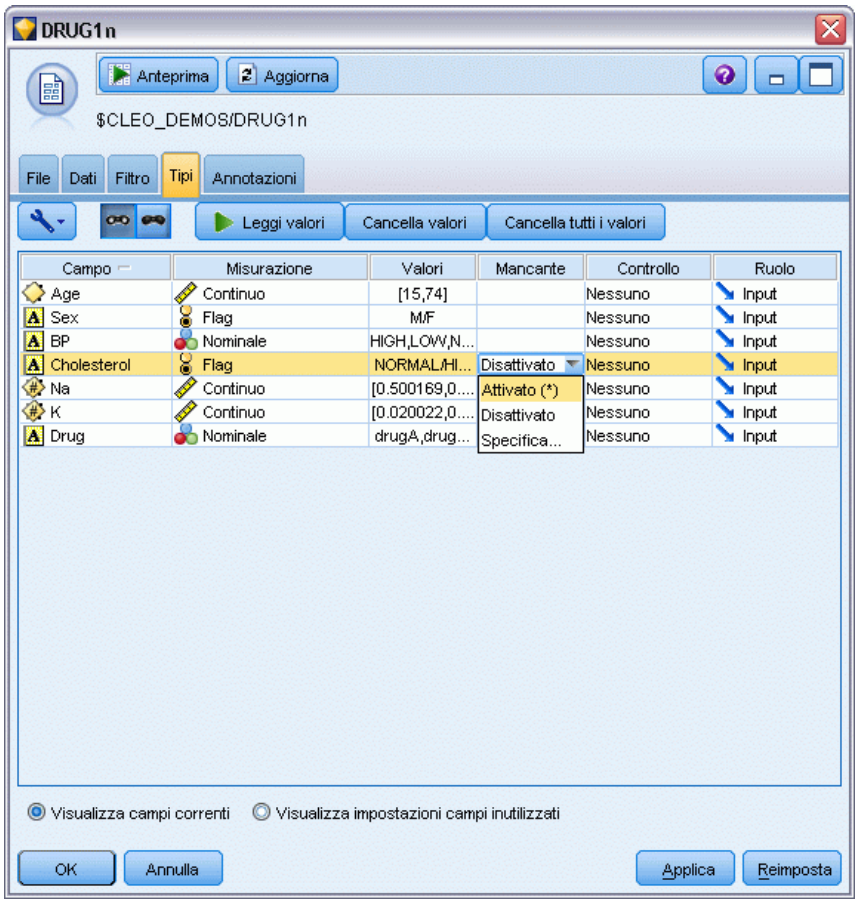


Figura 8-4
Selezione delle opzioni per i valori nella scheda Tipi



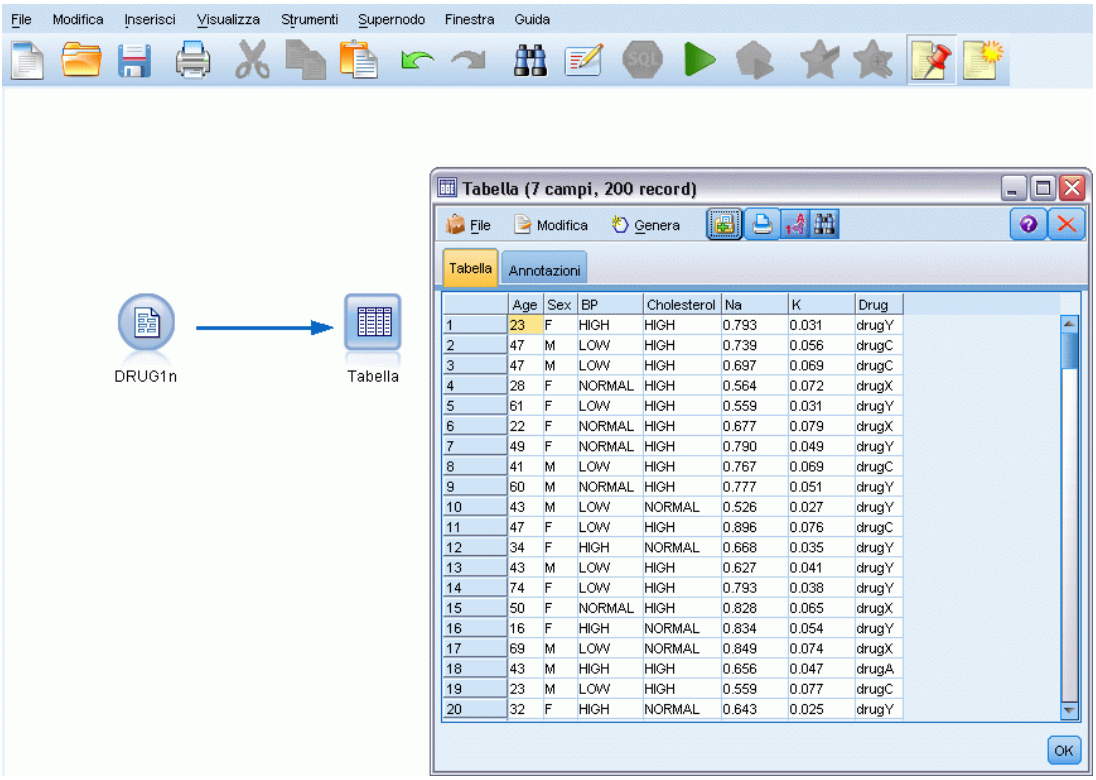
Fare clic sulla scheda Dati per sovrascrivere e modificare la **classe di archiviazione** di un campo. Si noti che la classe di archiviazione è diversa dalla **Misurazione**, ovvero il livello di misurazione (o tipo di utilizzo) del campo di dati. La scheda Tipi consente di ottenere ulteriori informazioni relative ai tipi di campi dei dati. È inoltre possibile fare clic su Leggi valori per visualizzare gli effettivi valori di ogni campo in relazione ai valori selezionati nella colonna *Valori*. Tale processo è anche denominato **istanziazione**.

Aggiunta di un nodo Tabella

Dopo avere caricato il file di dati, è possibile esaminare i valori di alcuni record. È per esempio possibile creare uno stream che includa un nodo Tabella. Per inserire un nodo Tabella nello stream, fare doppio clic sulla relativa icona nella palette oppure trascinarlo nell'area di disegno.

Figura 8-5
Nodo Tabella connesso alla sorgente dati

Figura 8-6
Esecuzione di uno stream dalla barra degli strumenti



	Age	Sex	BP	Cholesterol	Na	K	Drug
1	23	F	HIGH	HIGH	0.793	0.031	drugY
2	47	M	LOW	HIGH	0.739	0.056	drugC
3	47	M	LOW	HIGH	0.697	0.069	drugC
4	28	F	NORMAL	HIGH	0.564	0.072	drugX
5	61	F	LOW	HIGH	0.559	0.031	drugY
6	22	F	NORMAL	HIGH	0.677	0.079	drugX
7	49	F	NORMAL	HIGH	0.790	0.049	drugY
8	41	M	LOW	HIGH	0.767	0.069	drugC
9	60	M	NORMAL	HIGH	0.777	0.051	drugY
10	43	M	LOW	NORMAL	0.526	0.027	drugY
11	47	F	LOW	HIGH	0.896	0.076	drugC
12	34	F	HIGH	NORMAL	0.668	0.035	drugY
13	43	M	LOW	HIGH	0.627	0.041	drugY
14	74	F	LOW	HIGH	0.793	0.038	drugY
15	50	F	NORMAL	HIGH	0.828	0.065	drugX
16	16	F	HIGH	NORMAL	0.834	0.054	drugY
17	69	M	LOW	NORMAL	0.849	0.074	drugX
18	43	M	HIGH	HIGH	0.656	0.047	drugA
19	23	M	LOW	HIGH	0.559	0.077	drugC
20	32	F	HIGH	NORMAL	0.643	0.025	drugY

Facendo doppio clic su un nodo nella palette, tale nodo verrà automaticamente connesso a quello selezionato nell'area di disegno dello stream. In alternativa, qualora i nodi non siano già connessi, è possibile utilizzare il pulsante centrale del mouse per connettere il nodo Input al nodo Tabella. Per simulare il pulsante centrale del mouse, tenere premuto il tasto Alt mentre si utilizza il mouse. Per visualizzare la tabella, fare clic sul pulsante contrassegnato da una freccia verde sulla barra degli strumenti per eseguire lo stream, oppure fare clic con il pulsante destro del mouse sul nodo Tabella e scegliere Esegui.

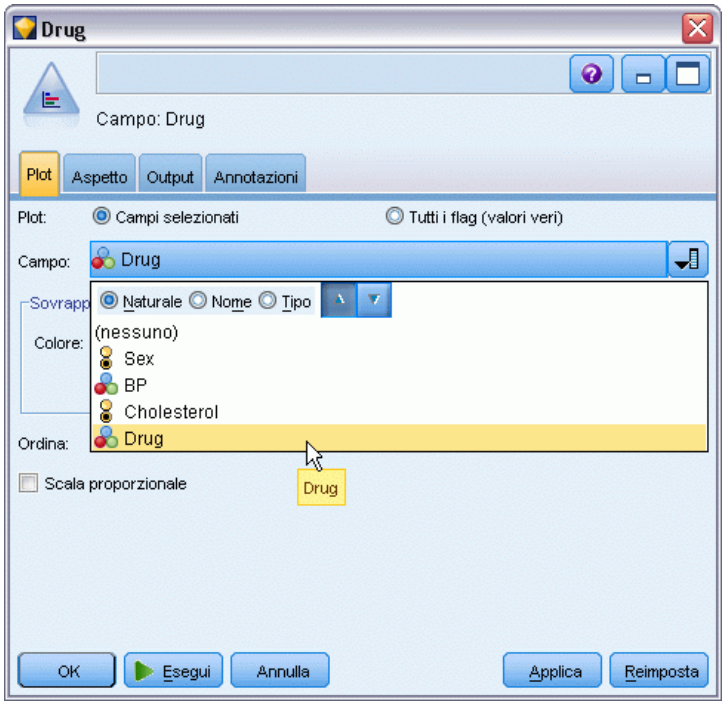
Creazione del grafico di un nodo Distribuzione

Durante il data mining, è spesso utile esplorare i dati creando riepiloghi visivi. IBM® SPSS® Modeler consente di scegliere tra diversi tipi di grafici, in base al tipo di dati che si desidera riepilogare. Per esempio, per individuare la proporzione della reazione dei pazienti alle singole cure, utilizzare un nodo Distribuzione.

Aggiungere un nodo Distribuzione allo stream e connetterlo al nodo Input, quindi fare doppio clic sul nodo per modificare le opzioni da visualizzare.

Selezionare *Cura* come campo obiettivo per il quale si desidera visualizzare la distribuzione. Fare quindi clic su Esegui nella finestra di dialogo.

Figura 8-7
Selezione di Cura come campo obiettivo



Il grafico risultante consente di visualizzare la "forma" dei dati. Viene dimostrato che i pazienti hanno reagito più alla cura *Y* che alle cure *B* e *C*.

Figura 8-8
Distribuzione della risposta al tipo di cura

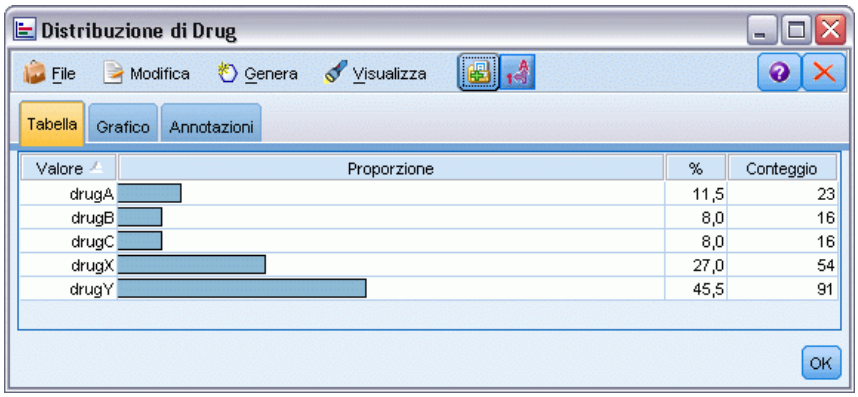
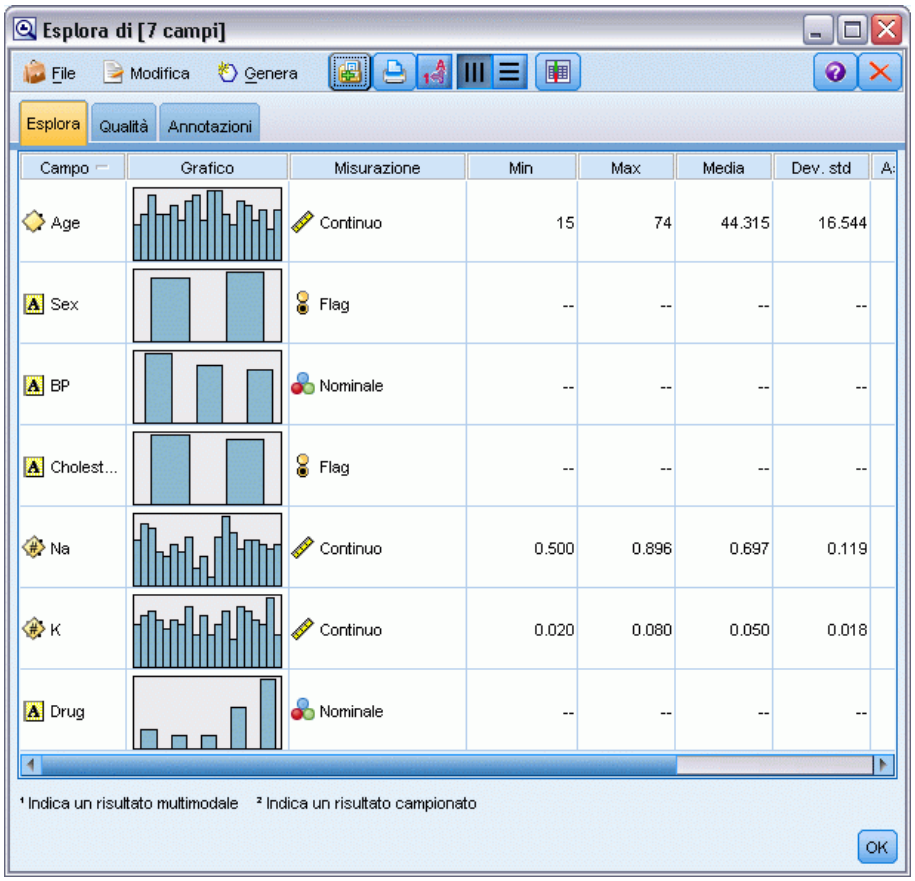


Figura 8-9
Risultati di un nodo Esplora



In alternativa, è possibile collegare ed eseguire un nodo Esplora per un rapido sguardo a distribuzioni e istogrammi per tutti i campi. Il nodo Esplora è disponibile nella scheda Output.

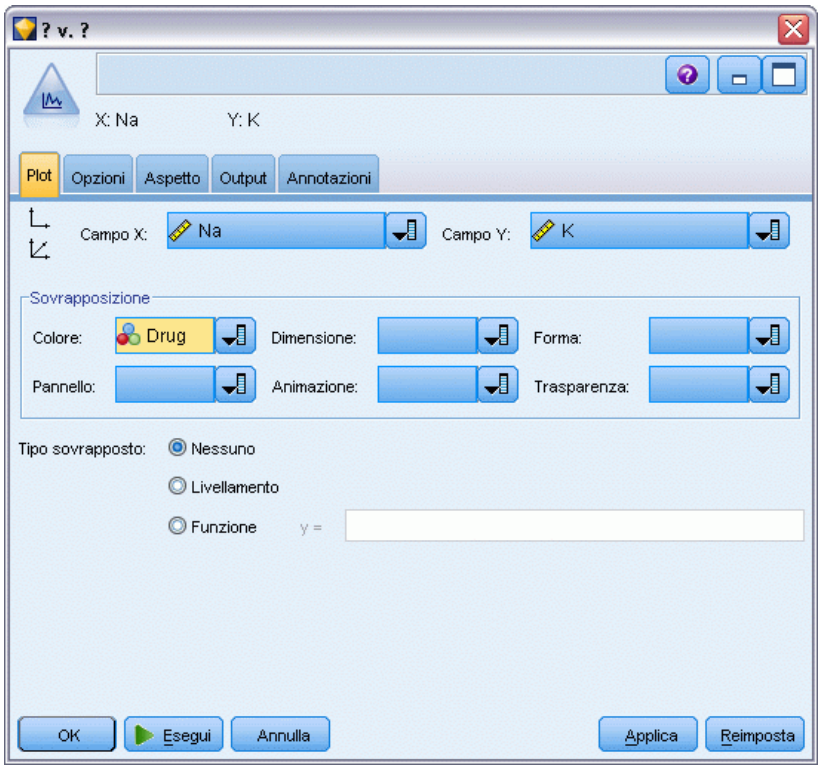
Creazione di un grafico a dispersione

È possibile analizzare ora i fattori che possono influenzare la variabile obiettivo *Cura*. I ricercatori sanno che le concentrazioni di sodio e potassio nel sangue rappresentano fattori importanti. Poiché si tratta di valori entrambi numerici, è possibile creare un grafico a dispersione per mettere a confronto sodio e potassio, utilizzando la categoria relativa alla cura come sovrapposizione di colore.

Posizionare un nodo Plot nell'area di lavoro, connetterlo al nodo Input e fare doppio clic per modificare il nodo.

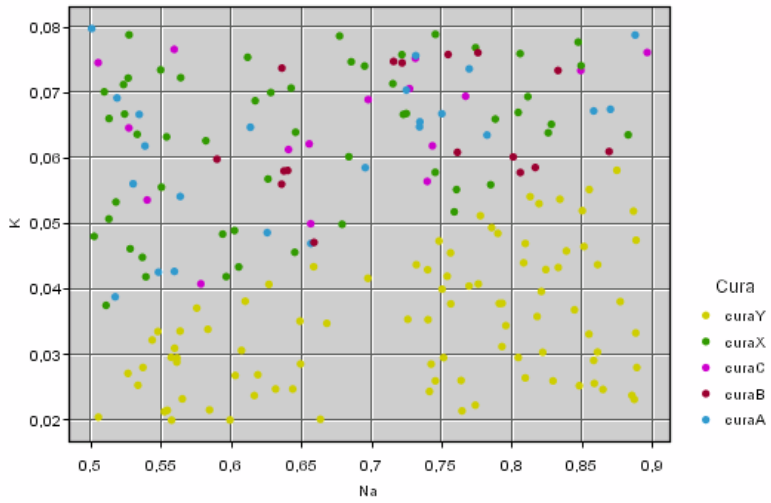
Nella scheda Plot, selezionare *Na* come campo X, *K* come campo Y e *Cura* come campo di sovrapposizione. Fare clic su Esegui.

Figura 8-10
Creazione di un grafico a dispersione



Nel grafico viene chiaramente mostrata una soglia sopra la quale la cura adatta è sempre la cura *Y* e sotto la quale la cura adatta non è mai la cura *Y*. Tale soglia rappresenta il rapporto del sodio (*Na*) rispetto al potassio (*K*).

Figura 8-11
Grafico a dispersione della distribuzione della cura

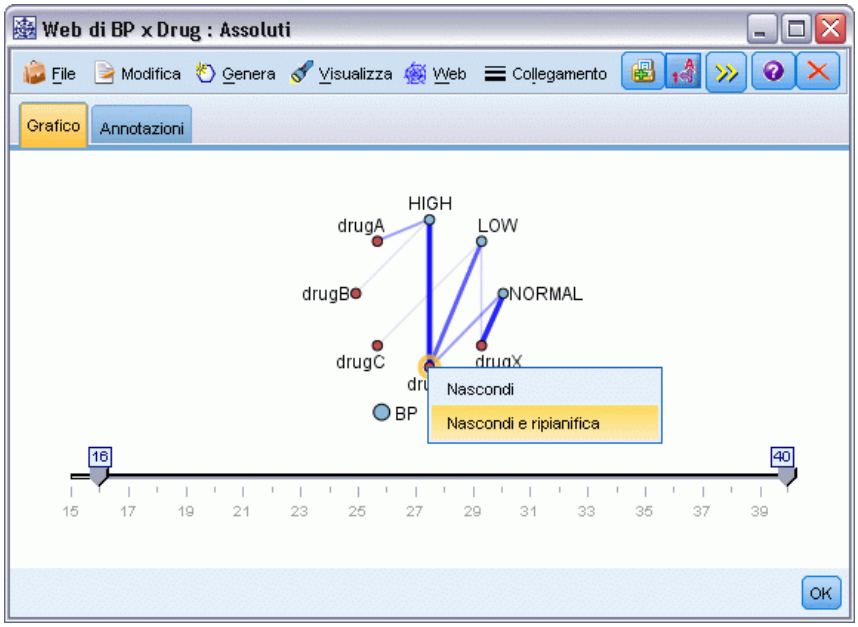


Creazione di un grafico Web

Poiché molti campi di dati sono categoriali, può essere utile rappresentare un grafico Web, che mappa le associazioni tra categorie diverse. Connettere innanzitutto un nodo Web a un nodo Input nell'area di lavoro. Nella finestra di dialogo del nodo Web, selezionare *Pressione* (per pressione sanguigna) e *Cura*. Fare clic su Esegui.

Dal plot, si noterà che la cura *Y* è associata a tutti e tre i livelli di pressione sanguigna. Non si tratta di un risultato inaspettato, in quanto è già stato determinato il caso in cui la cura *Y* risulta la più appropriata. Per prendere in esame le altre cure, è possibile nascondere la cura *Y*. Dal menu Visualizza, scegliere Modalità modifica, fare clic con il pulsante destro del mouse sul punto della cura *Y* e selezionare Nascondi e ripianifica.

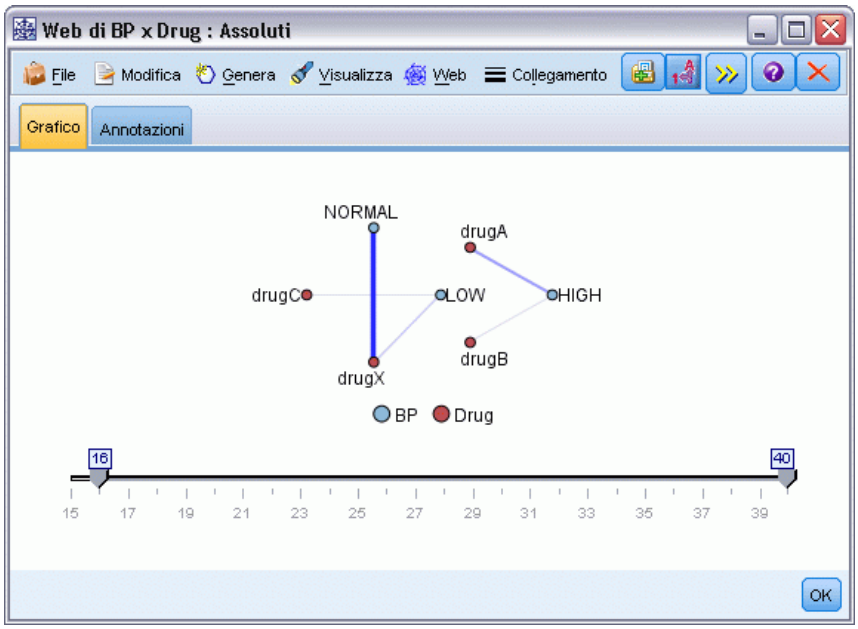
Figura 8-12
Grafico Web delle cure rispetto alla pressione sanguigna



Nel grafico semplificato, la cura *Y* e tutti i relativi collegamenti sono nascosti. È quindi possibile vedere chiaramente che alla pressione sanguigna alta sono associate solo le cure *A* e *B*, alla pressione sanguigna bassa sono associate solo le cure *C* e *X* e alla pressione sanguigna normale è associata solo la cura *X*. A questo punto non è tuttavia ancora possibile determinare come

effettuare una scelta tra le cure *A* e *B* e tra le cure *C* e *X* per un determinato paziente. In questo caso può essere utile la modellazione.

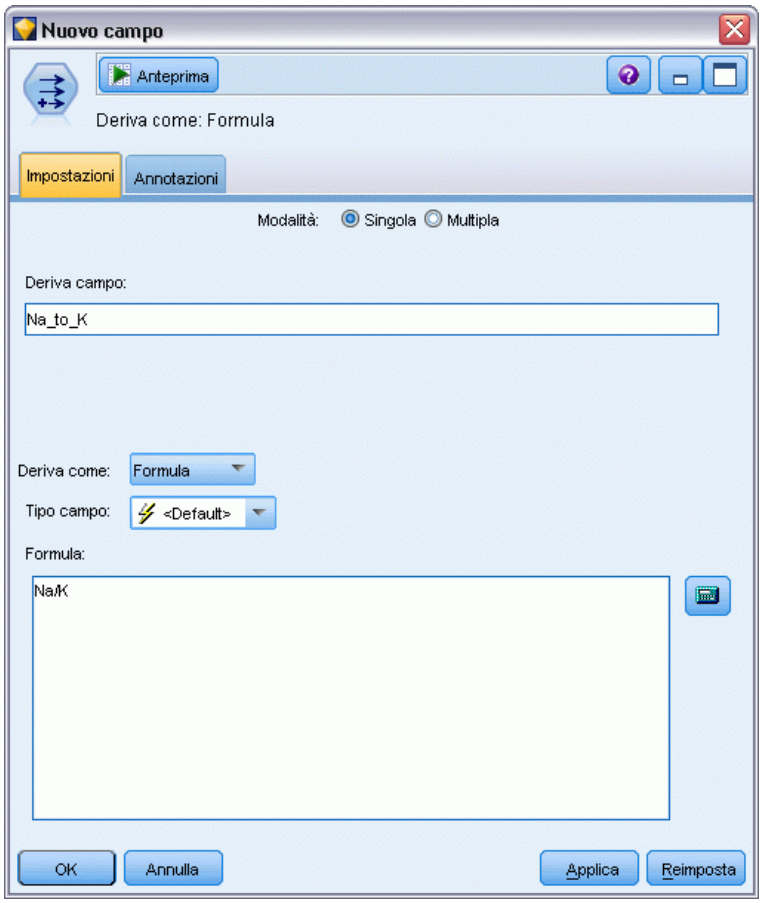
Figura 8-13
Grafico Web con cura Y e relativi collegamenti nascosti



Derivazione di un nuovo campo

Poiché il rapporto tra sodio e potassio sembra essere decisivo per l'individuazione dei casi in cui utilizzare la cura *Y*, è possibile derivare un campo contenente il valore di tale rapporto per ogni record. Tale campo sarà utile successivamente per la creazione di un modello che consenta di prevedere i casi in cui deve essere utilizzata ognuna delle cinque cure. Per semplificare il layout dello stream, iniziare a eliminare tutti i nodi tranne il nodo di input DRUG1n. Collegare un nodo Nuovo campo (scheda Oper su campi) a DRUG1n, quindi fare doppio clic sul nodo Nuovo campo per modificarlo.

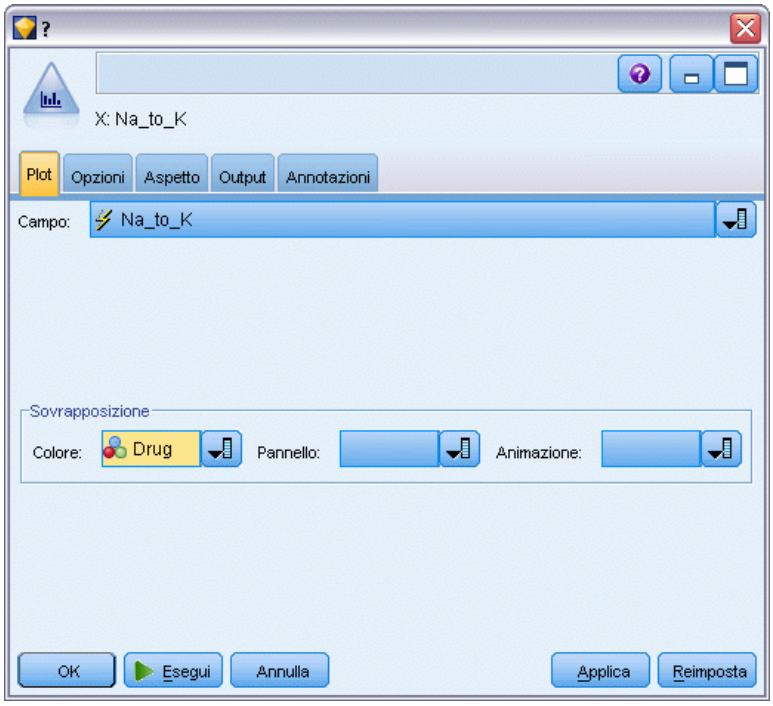
Figura 8-14
Modifica del nodo Nuovo campo



Denominare il nuovo campo *Na_su_K*. Poiché il nuovo campo è stato ottenuto dividendo il valore del sodio per quello del potassio, immettere Na/K nel campo della formula. È inoltre possibile creare una formula facendo clic sull'icona a destra del campo. Verrà visualizzato il generatore di espressioni che consente di creare in modo interattivo espressioni tramite elenchi incorporati di funzioni, operandi, campi e relativi valori.

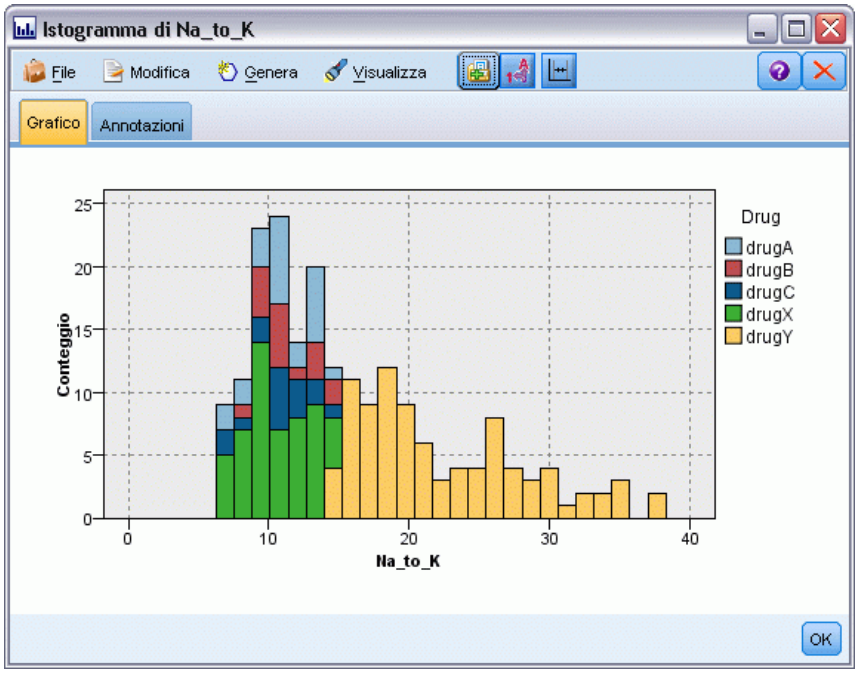
È possibile verificare la distribuzione del nuovo campo collegando un nodo Istogramma al nodo Nuovo campo. Nella finestra di dialogo del nodo Istogramma specificare *Na_su_K* come campo da rappresentare graficamente e *Cura* come campo di sovrapposizione.

Figura 8-15
Modifica del nodo Istogramma



Quando si esegue lo stream, viene visualizzato questo grafico. In base alla visualizzazione, sarà possibile concludere che quando il valore di *Na_su_K* è circa 15 o superiore, la cura appropriata è la cura *Y*.

Figura 8-16
Visualizzazione dell'istogramma

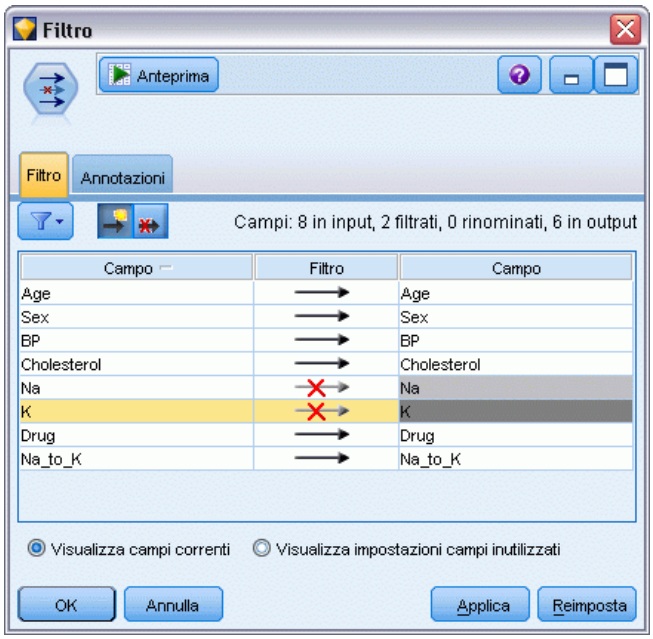


Creazione di un modello

Tramite l'esplorazione e la manipolazione dei dati, sono state formulate alcune ipotesi. Il rapporto del sodio rispetto al potassio nel sangue sembra influenzare la scelta della cura, così come la pressione sanguigna. Tuttavia, non è ancora possibile spiegare completamente tale relazione. In questo caso è utile ricorrere alla modellazione. Si tenterà di adattare i dati utilizzando il modello di creazione delle regole C5.0.

Poiché si utilizza un campo derivato, ovvero *Na_su_K*, è possibile filtrare i campi originali, *Na* e *K*, affinché non vengano utilizzati due volte nell'algoritmo di creazione di modelli. È possibile eseguire questa operazione utilizzando un nodo Filtro.

Figura 8-17
Modifica del nodo Filtro

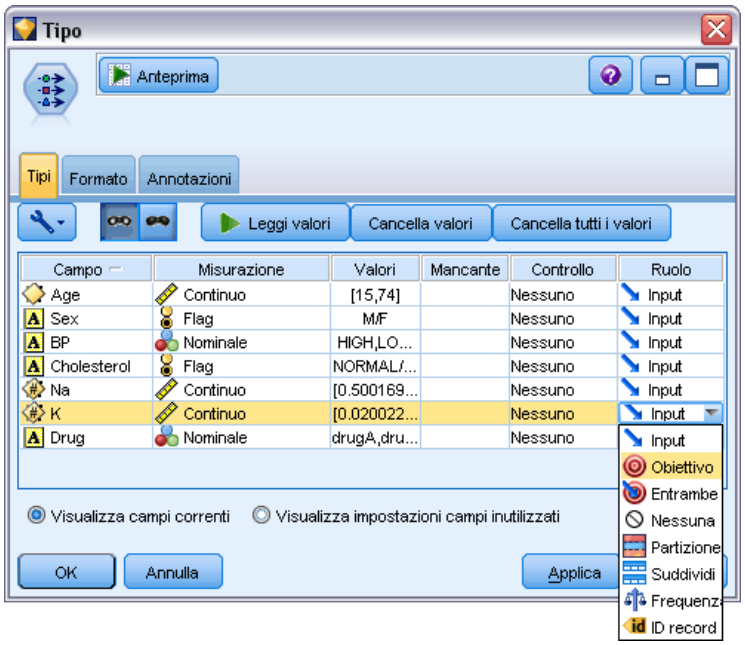


Nella scheda Filtro, fare clic sulle frecce accanto a *Na* e *K*. Verrà visualizzata una X rossa su ogni freccia, a indicare che i campi sono stati filtrati.

Quindi, collegare un nodo Tipo connesso a un nodo Filtro. Il nodo Tipo consente di indicare i tipi di campi che si stanno utilizzando e il modo in cui tali campi vengono utilizzati per prevedere i risultati.

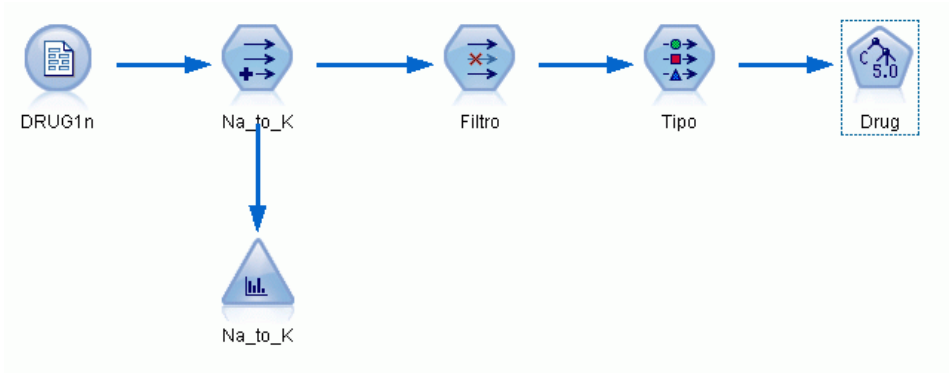
Sulla scheda Tipi, impostare il ruolo del campo *Cura* su Obiettivo, a indicare che *Cura* è il campo che si desidera prevedere. Lasciare il ruolo degli altri campi impostato su Input in modo che vengano utilizzati come predittori.

Figura 8-18
Modifica del nodo Tipo



Per stimare il modello, posizionare un nodo C5.0 nell'area di lavoro e collegarlo all'estremità dello stream, come illustrato. Fare quindi clic sul pulsante verde Esegui nella barra degli strumenti per eseguire lo stream.

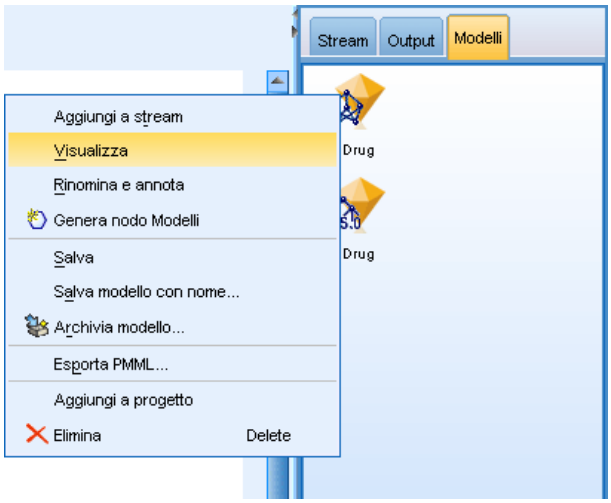
Figura 8-19
Aggiunta di un nodo C5.0



Visualizzazione del modello

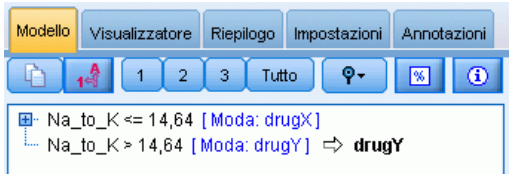
Quando si esegue il nodo C5.0, l'insieme di modelli viene aggiunto allo stream e alla palette Modelli nell'angolo superiore destro della finestra. Per visualizzare il modello, fare clic con il pulsante destro del mouse su una delle icone e scegliere Modifica o Visualizza dal menu di scelta rapida.

Figura 8-20
Visualizzazione del modello



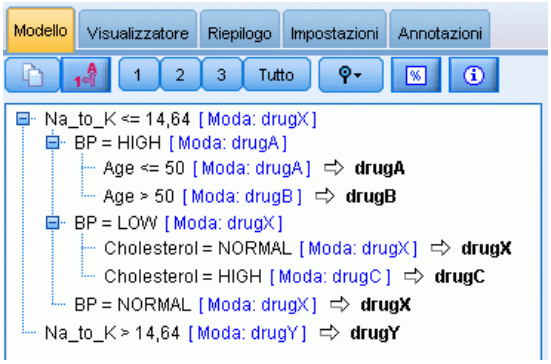
Tramite il browser relativo alla regola, verrà visualizzato l'insieme delle regole generate dal nodo C5.0 in formato di albero decisionale. Inizialmente, l'albero è compresso. Per espanderlo, fare clic sul pulsante Tutto per visualizzare tutti i livelli.

Figura 8-21
Browser delle regole



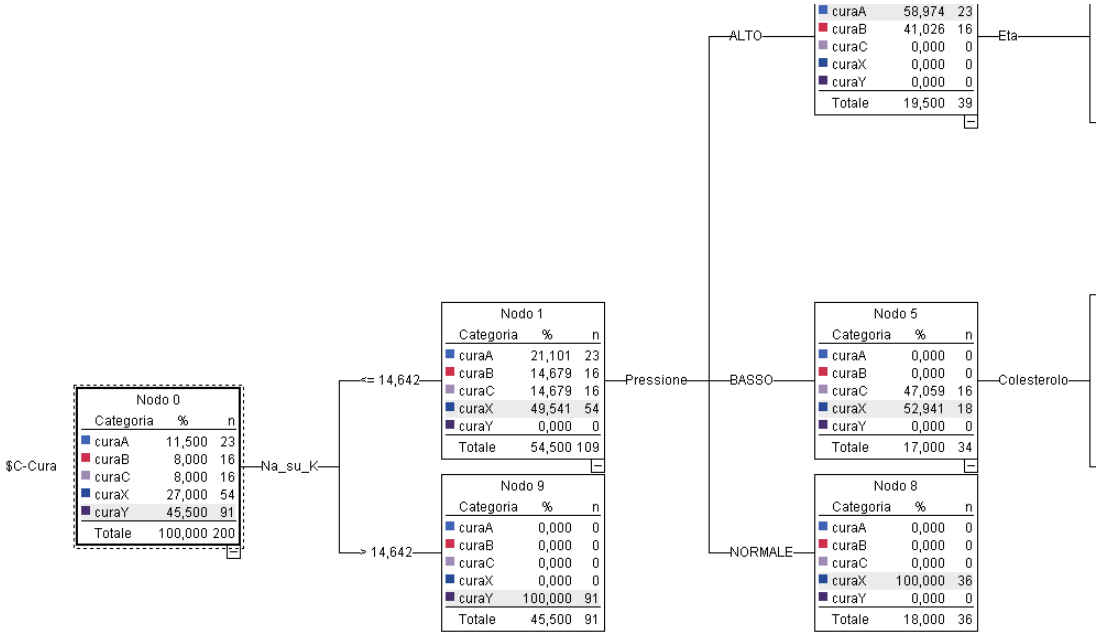
A questo punto è possibile ottenere le informazioni mancanti. Per i pazienti con un rapporto *Na*_su_*K* inferiore a 14,64 e pressione sanguigna alta, il fattore che determina la scelta della cura è l'età. Per le persone con pressione sanguigna bassa, il livello di colesterolo sembra essere il miglior predittore.

Figura 8-22
Browser delle regole completamente espanso



È possibile visualizzare lo stesso albero decisionale in un formato grafico più sofisticato selezionando la scheda Visualizzatore. In questa scheda è possibile visualizzare in modo più semplice il numero dei casi per ogni categoria relativa alla pressione sanguigna, nonché le percentuali dei casi.

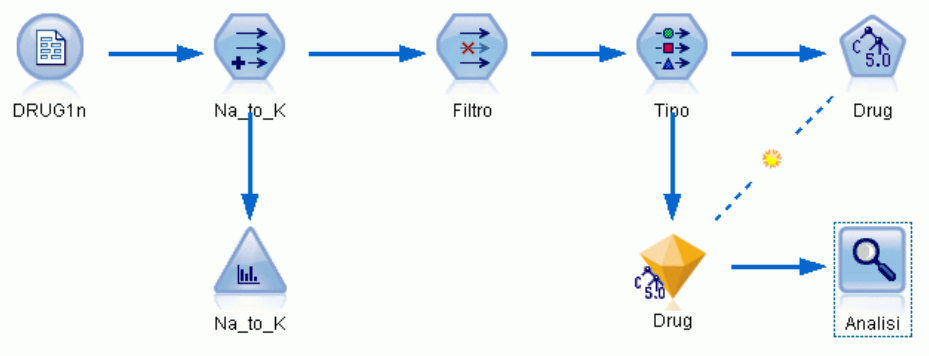
Figura 8-23
Albero decisionale in formato grafico



Utilizzo di un nodo Analisi

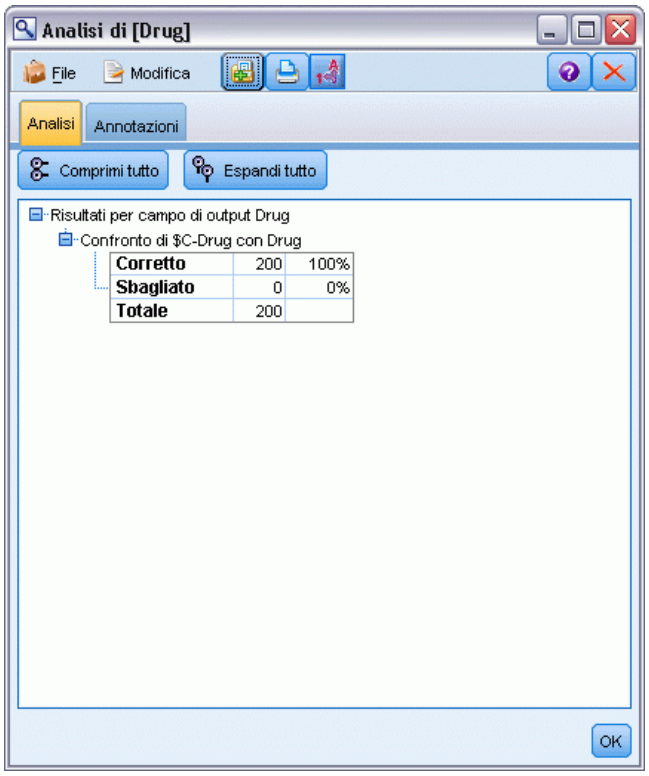
È possibile valutare la precisione del modello utilizzando un nodo Analisi. Collegare un nodo Analisi (dalla palette del nodo Output) all'insieme di modelli, aprire il nodo Analisi e fare clic su Esegui.

Figura 8-24
Aggiunta di un nodo Analisi



L'output del nodo Analisi dimostra che, con questo insieme di dati artificiale, il modello prevede correttamente la scelta della cura per tutti i record dell'insieme di dati. Con un insieme di dati reale, è poco probabile che si ottenga un grado di precisione pari al 100%, ma si potrà utilizzare il nodo Analisi per determinare se il modello è sufficientemente preciso per la propria applicazione.

Figura 8-25
Output del nodo Analisi



Capitolo

9

Screening dei predittori (Selezione funzioni)

Il nodo Selezione funzioni facilita l'identificazione dei campi più importanti per la previsione di un determinato risultato. Da un insieme di centinaia o addirittura migliaia di predittori, il nodo Selezione funzioni esegue lo screening, classifica e seleziona i predittori potenzialmente più importanti. Sostanzialmente, grazie a questo nodo è possibile ottenere un modello più veloce ed efficiente che utilizza un numero inferiore di predittori, di più rapida esecuzione e più semplice da capire.

I dati utilizzati in questo esempio rappresentano un data warehouse di un ipotetico gestore telefonico e contengono informazioni sulle adesioni a una speciale promozione da parte di 5000 clienti della società. I dati comprendono numerosi campi contenenti l'età, la professione, il reddito e le statistiche d'uso del telefono dei clienti. Tre campi "obiettivo" mostrano se il cliente ha aderito a ciascuna delle tre offerte che gli sono state proposte. La società desidera utilizzare questi dati per prevedere quali clienti hanno le maggiori probabilità di aderire a offerte analoghe in futuro.

In questo esempio viene utilizzato lo stream denominato *featureselection.str*, che fa riferimento al file di dati denominato *customer_dbase.sav*. Questi file sono disponibili nella directory *Demos* di tutte le installazioni di IBM® SPSS® Modeler. L'accesso si effettua dal gruppo di programmi IBM® SPSS® Modeler nel menu Start di Windows. Il file *featureselection.str* si trova nella directory *streams*.

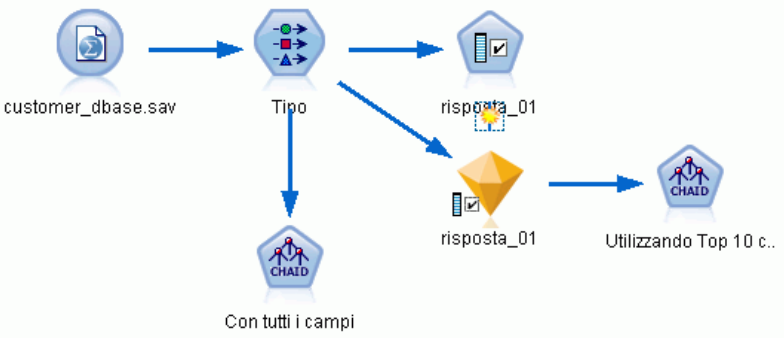
Questo esempio prende in esame come obiettivo solo una delle offerte, utilizzando il nodo CHAID di creazione dell'albero per sviluppare un modello in grado di descrivere il tipo di clienti con le maggiori probabilità di aderire alla promozione. Nell'esempio vengono messi a confronto due diversi modi di procedere:

- Senza selezione funzioni. Tutti i campi predittore dell'insieme di dati vengono utilizzati come input per l'albero CHAID.
- Con selezione funzioni. Viene utilizzato il nodo Selezione funzioni per selezionare i 10 predittori migliori che vengono poi immessi nell'albero CHAID.

Il confronto dei due modelli di alberi così ottenuti permette di verificare l'efficacia dei risultati ottenuti con la selezione delle funzioni.

Creazione dello stream

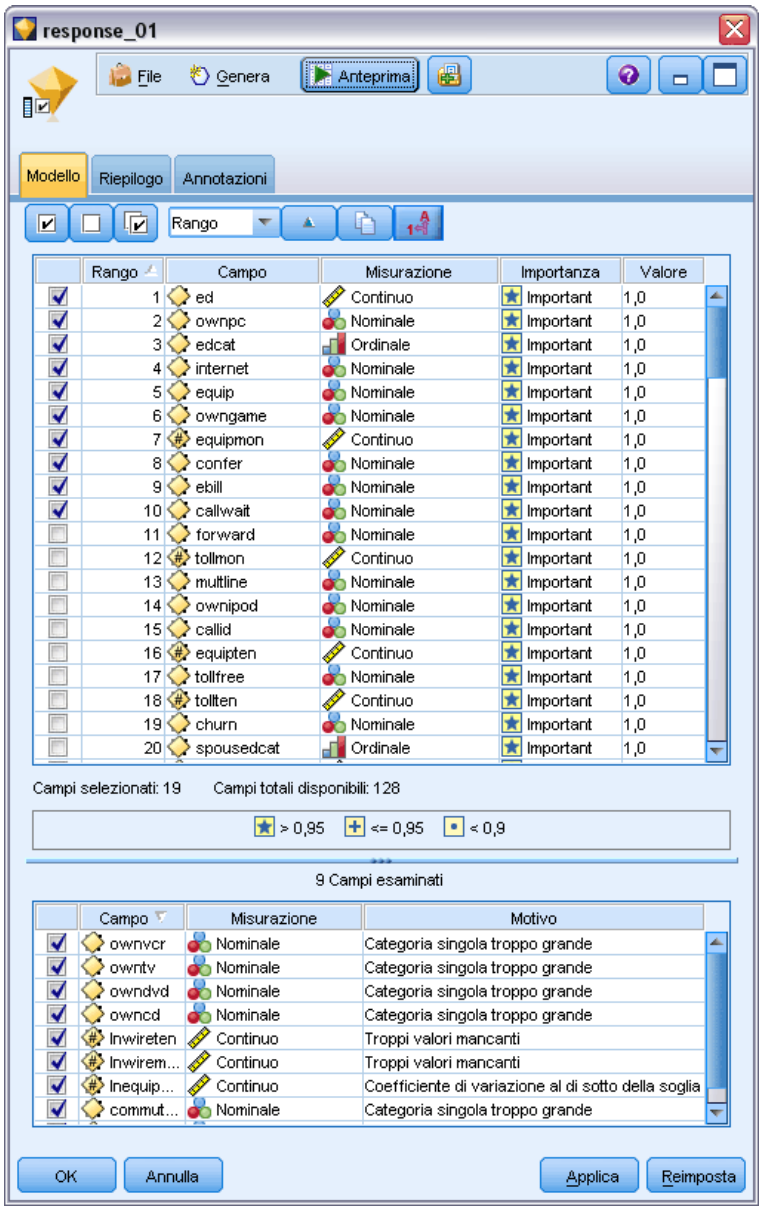
Figura 9-1
Esempio di stream con il nodo Selezione funzioni



- Collocare un nodo di input File Statistics in un'area di disegno dello stream vuota. Puntare il nodo al file di dati di esempio *customer_dbase.sav*, disponibile nella directory *Demos* dell'installazione di IBM® SPSS® Modeler in uso (in alternativa, aprire il file dello stream di esempio *featureselection.str* nella directory *streams*).
- Aggiungere un nodo Tipo. Nella scheda Tipi, scorrere verso il basso e cambiare il ruolo di *risposta_01* in *Obiettivo*. Impostare il ruolo su *Nessuno* per gli altri campi risposta (*risposta_02* e *risposta_03*) nonché per l'ID cliente (*IDclie*) in cima all'elenco. Per tutti gli altri campi, lasciare il ruolo impostato su *Input* e fare clic sul pulsante Leggi valori, quindi selezionare OK.
- Aggiungere un nodo Modelli Selezione funzioni allo stream. In questo nodo è possibile specificare le regole e i criteri per selezionare o escludere i campi.
- Eseguire lo stream per generare l'insieme di modelli Selezione funzioni.

- Con il pulsante destro del mouse, fare clic sull'insieme di modelli nello stream o nella palette Modelli e scegliere Modifica o Visualizza per visionare i risultati.

Figura 9-2
Scheda Modello nell'insieme di modelli Selezione funzioni



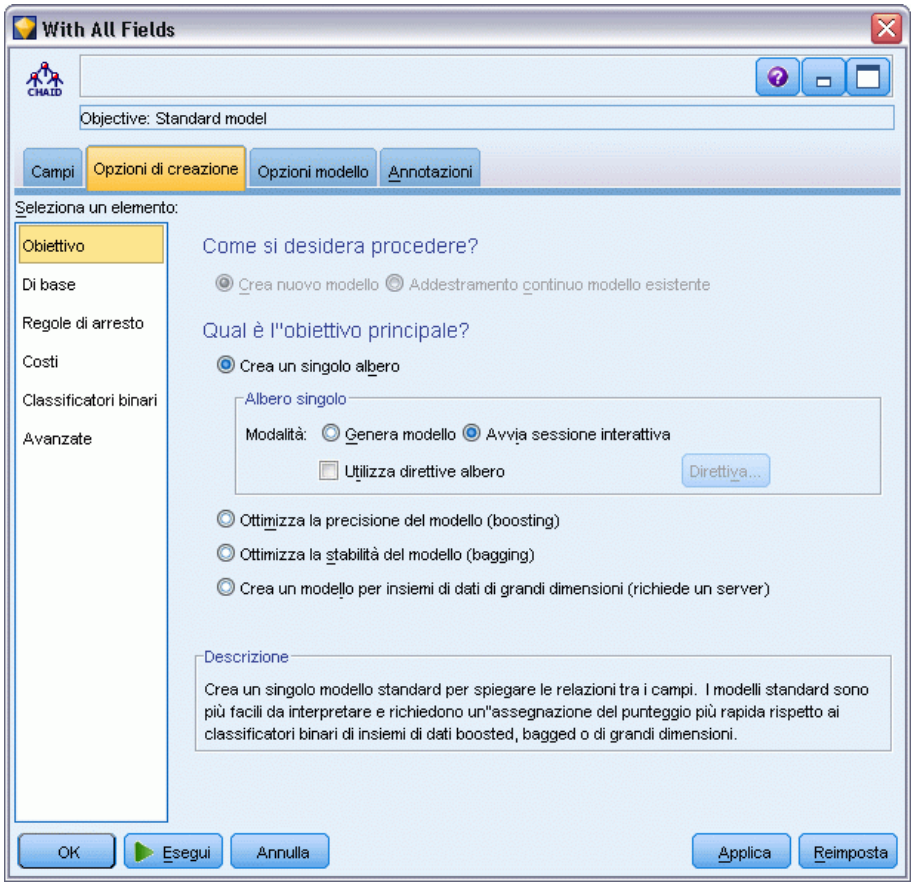
Il riquadro in alto mostra i campi che si sono rivelati utili per la previsione, classificati in ordine di importanza. Il riquadro in basso mostra i campi che sono stati esclusi dall'analisi e il motivo dell'esclusione. Esaminando i campi del riquadro in alto è possibile decidere quali usare nelle successive sessioni di modellazione.

- A questo punto è possibile selezionare i campi da utilizzare a valle. Benché in origine siano stati individuati 34 campi importanti, si desidera ridurre ulteriormente l'insieme dei predittori.

- Selezionare solo i 10 predittori principali utilizzando i segni di spunta nella prima colonna per deselezionare i predittori da escludere. (Fare clic sul segno di spunta nella riga 11, tenere premuto il tasto Maiusc e fare clic sul segno di spunta nella riga 34.) Chiudere l'insieme di modelli.
- Per confrontare i risultati senza eseguire una selezione funzioni, è necessario aggiungere due nodi di modellazione CHAID allo stream: uno che utilizza la Selezione funzioni e uno che non la utilizza.
- Collegare un nodo CHAID al nodo Tipo e l'altro all'insieme di modelli Selezione funzioni.
- Aprire ciascun nodo CHAID, selezionare la scheda Opzioni di creazione e verificare che le opzioni Crea nuovo modello, Crea un singolo albero e Avvia sessione interattiva siano selezionate nella scheda Obiettivi.

  Nel riquadro Di base, verificare che l'opzione Profondità massima albero sia impostata su 5.

Figura 9-3
Impostazioni degli obiettivi del nodo Modelli CHAID per tutti i campi predittore

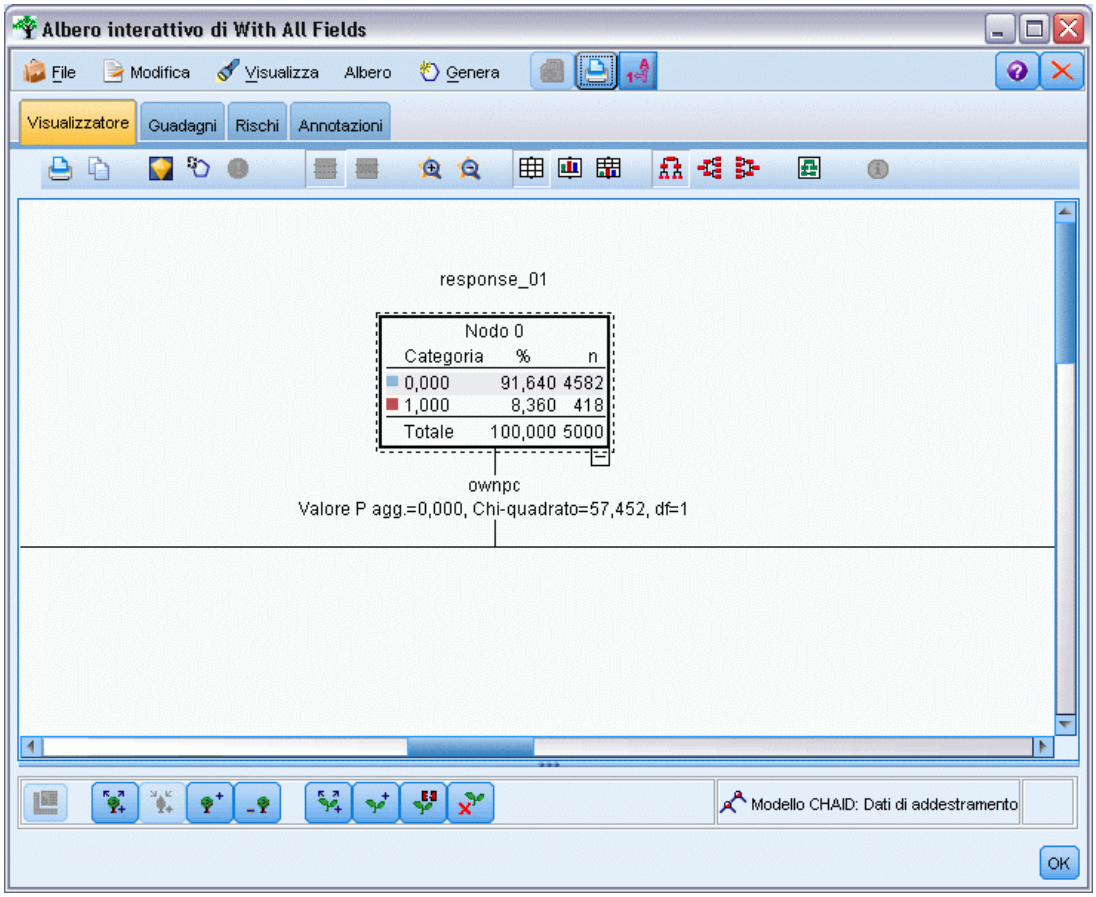


Creazione dei modelli

- Eseguire il nodo CHAID che utilizza tutti i predittori dell'insieme di dati (quello collegato al nodo Tipo). Durante l'esecuzione, osservare il tempo impiegato per portare a termine l'operazione. Nella finestra dei risultati viene visualizzata una tabella.

- Dai menu, scegliere Albero > Ingrandisci albero per ingrandire e visualizzare l'albero espanso.

Figura 9-4
Ingrandimento dell'albero nel Generatore alberi



- Ora ripetere le stesse operazioni per l'altro nodo CHAID, che utilizza solo 10 predittori. Anche in questo caso, espandere l'albero all'apertura del Generatore alberi.

Il secondo modello dovrebbe essere stato eseguito più rapidamente del primo. Poiché questo insieme di dati è relativamente piccolo, la differenza nei tempi di esecuzione è probabilmente di pochi secondi; tuttavia, con insiemi di dati più grandi che si incontrano nelle applicazioni reali, la differenza può essere considerevole, dell'ordine di minuti o addirittura ore. L'utilizzo della selezione funzioni può ridurre sensibilmente i tempi di elaborazione.

Inoltre, il secondo albero contiene un minor numero di nodi rispetto al primo ed è più facilmente interpretabile. Prima di decidere se utilizzarlo o meno, però, è necessario determinare se è efficace e quali sono le differenze rispetto al modello che utilizza tutti i predittori.

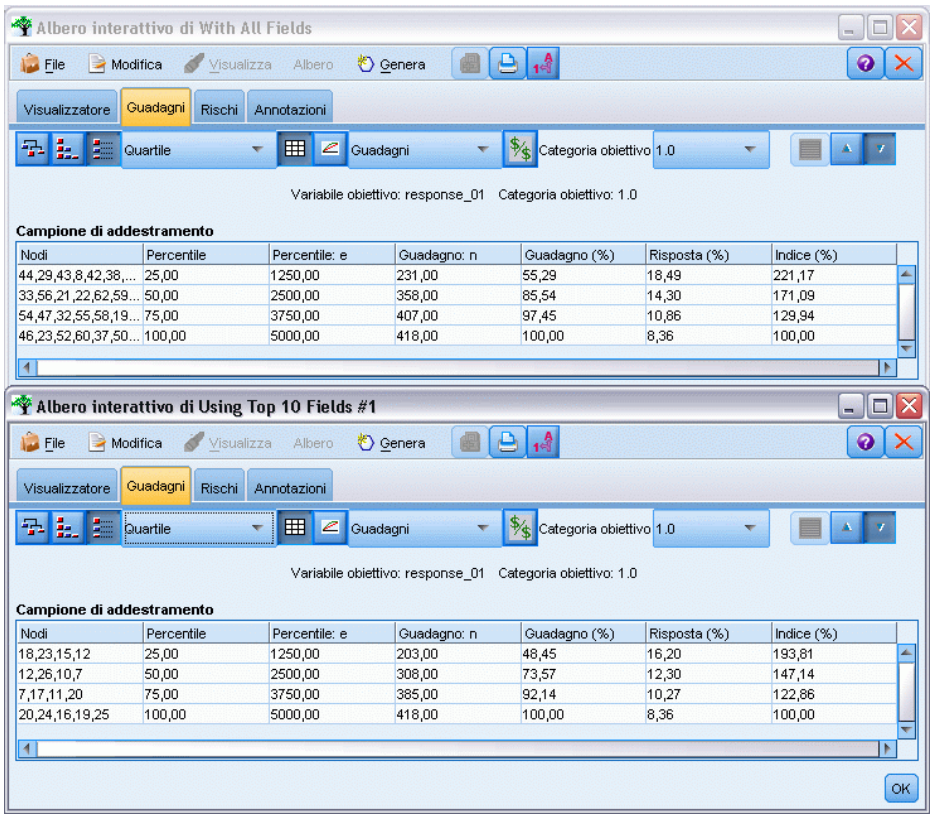
Confronto tra i risultati

Per confrontare i due risultati occorre un parametro di misura dell'efficacia. A tale scopo è possibile utilizzare la scheda Guadagni del Generatore alberi osservando il valore **lift**, che misura la probabilità dei record di un nodo di rientrare nella categoria obiettivo rispetto a tutti i record dell'insieme di dati. Per esempio, un valore di lift del 148% indica che i record del nodo hanno 1,48 probabilità in più di rientrare nella categoria obiettivo rispetto a tutti gli altri record dell'insieme di dati. Il Lift è indicato nella colonna *Indice* della scheda Guadagni.

- Fare clic sulla scheda Guadagni nel Generatore alberi dell'insieme completo dei predittori. Impostare la categoria obiettivo su 1.0. Impostare la visualizzazione sui quartili facendo clic prima sul pulsante della barra degli strumenti Quantili e quindi selezionando Quartile dall'elenco a discesa a destra di questo pulsante.

- Ripetere il procedimento nel Generatore alberi relativo all'insieme di 10 predittori in modo da ottenere due tabelle Guadagni simili da confrontare, come illustrato nelle figure seguenti.

Figura 9-5
Grafici dei guadagni relativi ai due modelli CHAID



Ciascuna tabella Guadagni raggruppa i nodi terminali del relativo albero in quartili. Per confrontare l'efficacia dei due modelli, osservare il lift (il valore *Indice*) del primo quartile di ogni tabella.

Se si includono tutti i predittori, il modello mostra un lift del 221%. Questo significa che i casi con le caratteristiche di questi nodi hanno 2,2 volte più probabilità di aderire alla promozione obiettivo. Per vedere quali sono queste caratteristiche, fare clic sulla prima riga per selezionarla e passare quindi alla scheda Visualizzatore, dove i nodi corrispondenti sono ora evidenziati in nero. Seguire l'albero fino ai singoli nodi terminali evidenziati per vedere come sono stati suddivisi i predittori. Il primo quartile comprende da solo 10 nodi. Se si trasporta il risultato nei modelli di calcolo del punteggio delle applicazioni reali, 10 profili cliente diversi possono essere difficili da gestire.

Se si includono solo i primi 10 predittori (identificati dalla selezione funzioni), il lift è circa il 194%. Sebbene non sia preciso quanto quello che utilizza tutti i predittori, questo modello è sicuramente utile. In questo caso, il primo quartile comprende solo quattro nodi ed è quindi più semplice. Pertanto, è possibile concludere che il Modello di selezione funzioni è preferibile rispetto a quello che utilizza tutti i predittori.

Riepilogo

Vengono riesaminati ora i vantaggi della selezione funzioni. L'utilizzo di un minor numero di predittori è meno costoso poiché la mole di dati da raccogliere, elaborare e inserire nei modelli è più contenuta. Anche il tempo di calcolo viene ridotto. In questo esempio, anche includendo la fase aggiuntiva della selezione funzioni, la creazione del modello è stata decisamente più rapida con l'utilizzo di un insieme di predittori più piccolo. Con insiemi di dati più cospicui utilizzati nelle applicazioni reali, il risparmio di tempo sarebbe notevolmente amplificato.

L'impiego di un minor numero di predittori determina una semplificazione nel calcolo dei punteggi. Come mostra l'esempio, è possibile identificare solo quattro profili di clienti che hanno una buona probabilità di aderire alla promozione. Si noti che con numeri più elevati di predittori si corre il rischio di sovradattare il modello. Il modello più semplice può essere generalizzato meglio per adattarlo ad altri insiemi di dati (anche se è necessario fare delle prove per esserne sicuri).

Per la selezione delle funzioni si sarebbe potuto utilizzare un algoritmo per la creazione di alberi, lasciando che fosse l'albero a individuare automaticamente i predittori più importanti. In effetti, l'algoritmo CHAID è utilizzato spesso a questo scopo, ed è anche possibile ingrandire l'albero un livello per volta per controllarne la profondità e la complessità. Tuttavia, il nodo Selezione funzioni è più veloce e più facile da usare. Esso classifica rapidamente tutti i predittori in un'unica operazione e consente di individuare velocemente i campi più importanti, nonché di variare il numero di predittori da includere. Questo esempio si potrebbe facilmente riprodurre utilizzando i primi 15 o 20 predittori anziché i primi 10, confrontando i risultati per determinare quale sia il modello ottimale.

Capitolo
10

Riduzione della lunghezza della stringa dei dati di input (Ricodifica)

Per i modelli di regressione logistica binomiale e classificatore automatico che includono un modello di regressione logistica binomiale, i campi stringa sono limitati a un numero massimo di otto caratteri. Le stringhe superiori a otto caratteri possono essere ricodificate mediante il nodo Ricodifica.

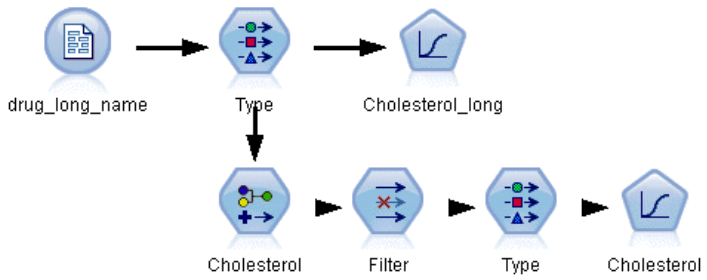
In questo esempio viene utilizzato lo stream denominato *reclassify_strings.str*, che fa riferimento al file di dati denominato *drug_long_name*. Questi file sono disponibili nella directory *Demos* di tutte le installazioni di IBM® SPSS® Modeler. L'accesso si effettua dal gruppo di programmi IBM® SPSS® Modeler nel menu Start di Windows. Il file *reclassify_strings.str* si trova nella directory *streams*.

Questo esempio è incentrato su una piccola parte di uno stream per mostrare il tipo di errori che possono essere generati in caso di stringhe troppo lunghe e spiega come utilizzare il nodo Ricodifica per modificare i dettagli della stringa portandola a una lunghezza accettabile. Sebbene l'esempio utilizzi un nodo Regressione logistica binomiale, è applicabile anche in caso di utilizzo del nodo Classificatore automatico per generare un modello di regressione logistica binomiale.

Ricodifica dei dati

- Utilizzando un nodo di input Testo variabile, collegarsi all'insieme di dati *drug_long_name* nella cartella *Demos*.

Figura 10-1
Stream di esempio che mostra la ricodifica della stringa per la regressione logistica binomiale



- Aggiungere un nodo Tipo al nodo Input e selezionare Colesterolo_lunga come obiettivo.
- Aggiungere un nodo Regressione logistica al nodo Tipo.

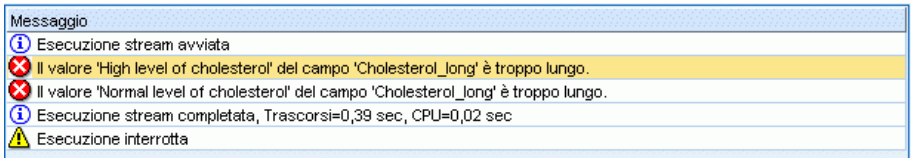
- Nel nodo Regressione logistica, fare clic sulla scheda Modello e selezionare la procedura Binomiale.

Figura 10-2
Dettagli della stringa lunga nel campo "Colesterolo_lunga"

- Quando si esegue il nodo Regressione logistica in *reclassify_strings.str*, viene visualizzato un messaggio di errore che informa che i valori della stringa Colesterolo_lunga sono troppo lunghi.

  Se si incontra questo tipo di messaggio di errore, attenersi alla procedura illustrata nel seguito di questo esempio per modificare i dati.

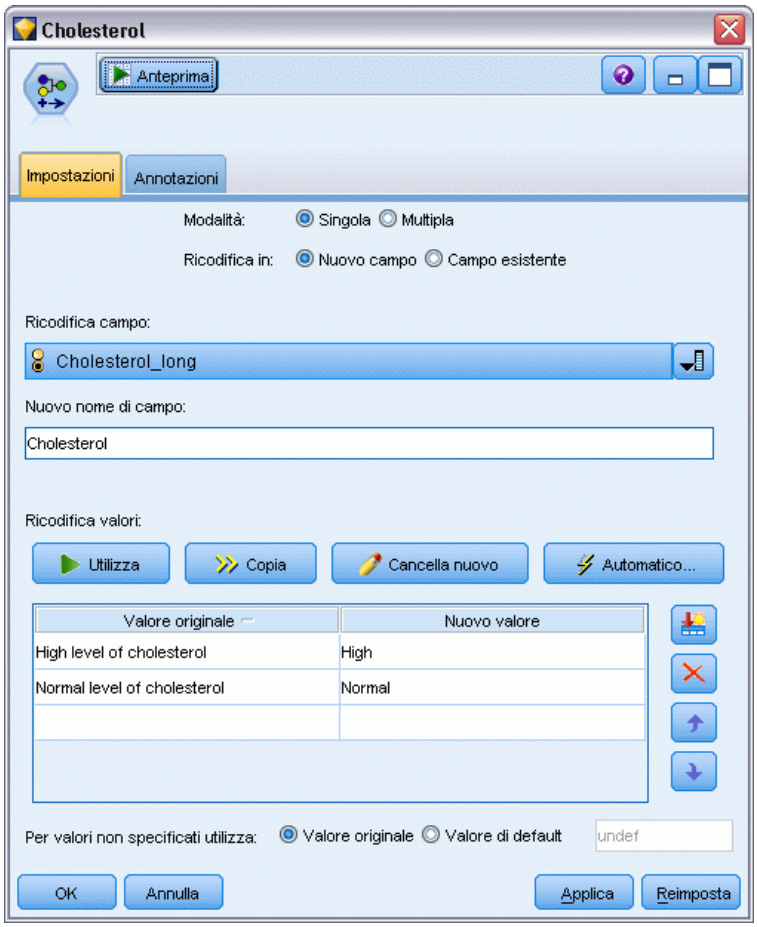
Figura 10-3
Messaggio di errore visualizzato durante l'esecuzione nel nodo di regressione logistica binomiale



- Aggiungere un nodo Ricodifica al nodo Tipo.
- Nel campo Ricodifica, selezionare Colesterolo_lunga.
- Immettere Colesterolo come nome del nuovo campo.
- Fare clic sul pulsante Recupera per aggiungere i valori Colesterolo_lunga alla colonna del valore originale.

- Nella colonna del nuovo valore, immettere Alto accanto al valore originale di Alto livello di colesterolo e Normale accanto al valore originale di Livello normale di colesterolo.

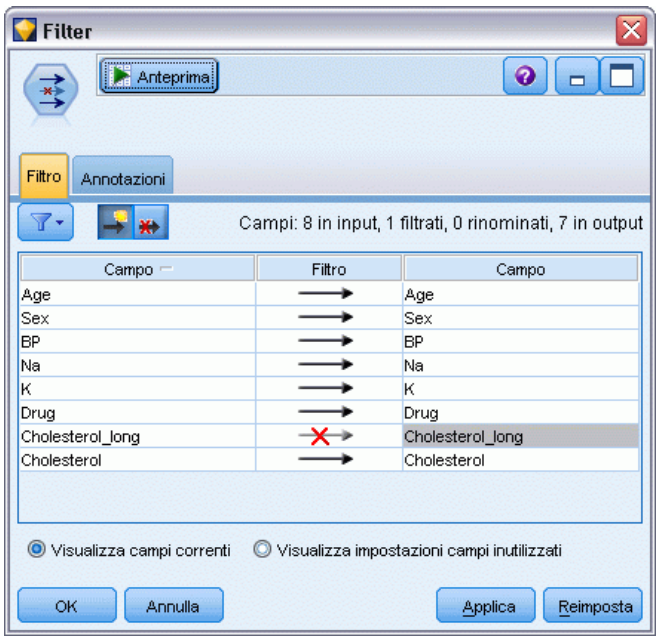
Figura 10-4
Ricodifica delle stringhe lunghe



- Aggiungere un nodo Filtro al nodo Ricodifica.

▶ Nella colonna Filtro, fare clic per eliminare Colesterolo_lunga.

Figura 10-5
Filtro del campo "Colesterolo_lunga" dai dati



▶ Aggiungere un nodo Tipo al nodo Filtro e selezionare Colesterolo come obiettivo.

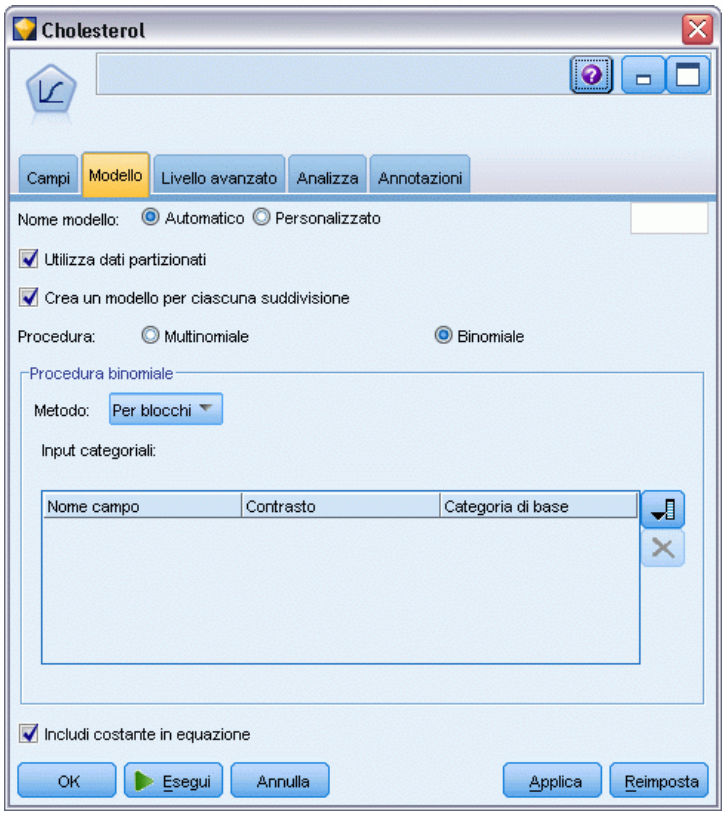
Figura 10-6
Dettagli della stringa corta nel campo "Colesterolo"

▶ Aggiungere un nodo Logistica al nodo Tipo.

▶ Nel nodo Logistica, fare clic sulla scheda Modello e selezionare la procedura Binomiale.

▶ È ora possibile eseguire il nodo Logistica binomiale e generare un modello senza visualizzare un messaggio di errore.

Figura 10-7
Scelta di Binomiale come procedura



Questo esempio illustra solo una parte di stream. Per ulteriori informazioni sui tipi di stream nei quali potrebbe essere necessario ricodificare le stringhe lunghe, sono disponibili i seguenti esempi:

- Nodo Classificatore automatico. Per ulteriori informazioni, vedere l'argomento Modellazione della risposta dei clienti (Classificatore automatico) in il capitolo 4 a pag. 45.
- nodo Regressione logistica binomiale Per ulteriori informazioni, vedere l'argomento Tasso di abbandono nelle telecomunicazioni (Regressione logistica binomiale) in il capitolo 13 a pag. 160.

Ulteriori informazioni sull'utilizzo di IBM® SPSS® Modeler, quali manuale dell'utente, riferimenti ai nodi e il manuale Algorithms Guide sono disponibili nella directory *\Documentation* del disco di installazione.

Parte III:
Esempi di modellazione

Capitolo

11

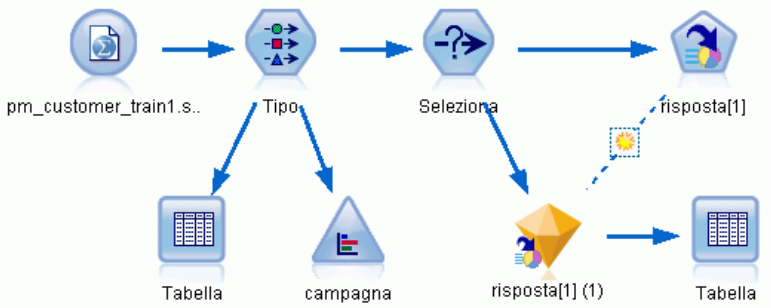
Modellazione della risposta dei clienti (Elenco decisionale)

L'algoritmo Elenco decisionale consente di generare delle regole che indicano una probabilità superiore o inferiore che si verifichi un determinato risultato binario (sì o no). I modelli Elenco decisionale sono ampiamente utilizzati nelle attività di CRM (Customer Relationship Management), come le applicazioni di call center o marketing.

Questo esempio è basato su una società fittizia che desidera ottenere risultati più redditizi nelle campagne di marketing future, inviando offerte mirate ai singoli clienti Specificatamente, l'esempio utilizza un modello Elenco decisionale per identificare le caratteristiche dei clienti che, stando ai risultati delle campagne precedenti, hanno più probabilità di rispondere in modo favorevole nonché, in base ai risultati ottenuti, generare una mailing list.

I modelli Elenco decisionale sono particolarmente adatti alla modellazione interattiva, poiché consentono di regolare i parametri del modello e di visualizzare immediatamente i risultati. Per un approccio diverso che consenta di creare automaticamente vari modelli diversi e di classificare i risultati, è possibile utilizzare invece il nodo Classificatore automatico.

Figura 11-1
Stream campione Elenco decisionale

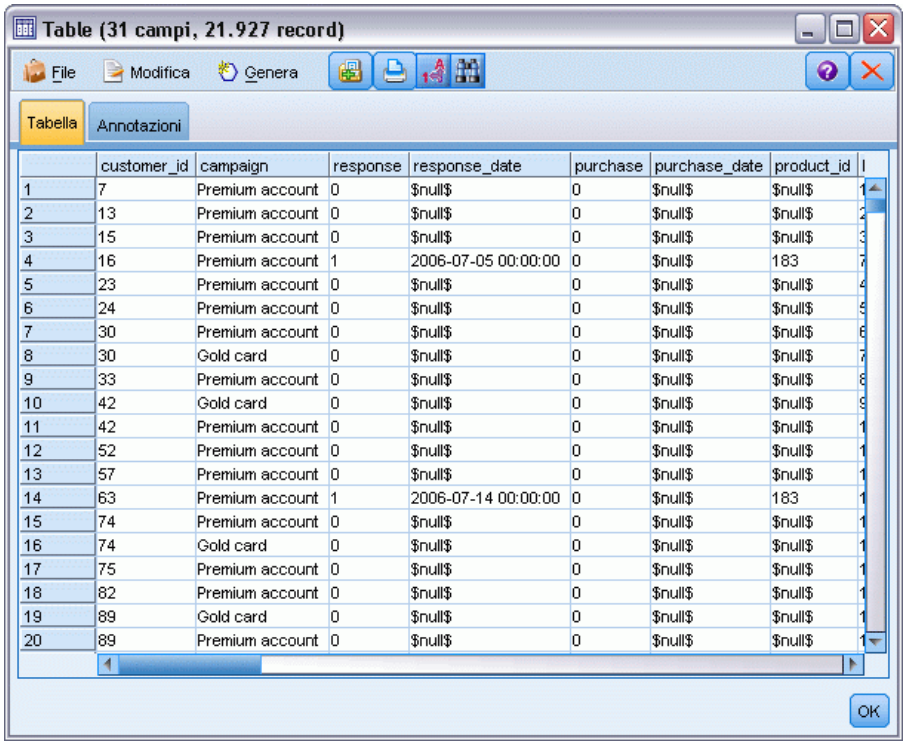


Questo esempio utilizza lo stream *pm_decisionlist.str*, che fa riferimento al file di dati *pm_customer_train1.sav*. Questi file sono disponibili nella directory *Demos* di tutte le installazioni di IBM® SPSS® Modeler. L'accesso si effettua dal gruppo di programmi IBM® SPSS® Modeler nel menu Start di Windows. Il file *pm_decisionlist.str* si trova nella directory *streams*.

Dati storici

Il file *pm_customer_train1.sav* include i dati storici che tengono traccia delle offerte fatte a specifici clienti nell'ambito delle campagne precedenti, come indicato dal campo *campagna*. Il numero maggiore di record rientra nella campagna *Premium account*.

Figura 11-2
Dati sulle promozioni precedenti



	customer_id	campaign	response	response_date	purchase	purchase_date	product_id	l
1	7	Premium account	0	$null$	0	$null$	$null$	1
2	13	Premium account	0	$null$	0	$null$	$null$	2
3	15	Premium account	0	$null$	0	$null$	$null$	3
4	16	Premium account	1	2006-07-05 00:00:00	0	$null$	183	7
5	23	Premium account	0	$null$	0	$null$	$null$	4
6	24	Premium account	0	$null$	0	$null$	$null$	5
7	30	Premium account	0	$null$	0	$null$	$null$	6
8	30	Gold card	0	$null$	0	$null$	$null$	7
9	33	Premium account	0	$null$	0	$null$	$null$	8
10	42	Gold card	0	$null$	0	$null$	$null$	9
11	42	Premium account	0	$null$	0	$null$	$null$	1
12	52	Premium account	0	$null$	0	$null$	$null$	1
13	57	Premium account	0	$null$	0	$null$	$null$	1
14	63	Premium account	1	2006-07-14 00:00:00	0	$null$	183	1
15	74	Premium account	0	$null$	0	$null$	$null$	1
16	74	Gold card	0	$null$	0	$null$	$null$	1
17	75	Premium account	0	$null$	0	$null$	$null$	1
18	82	Premium account	0	$null$	0	$null$	$null$	1
19	89	Gold card	0	$null$	0	$null$	$null$	1
20	89	Premium account	0	$null$	0	$null$	$null$	1

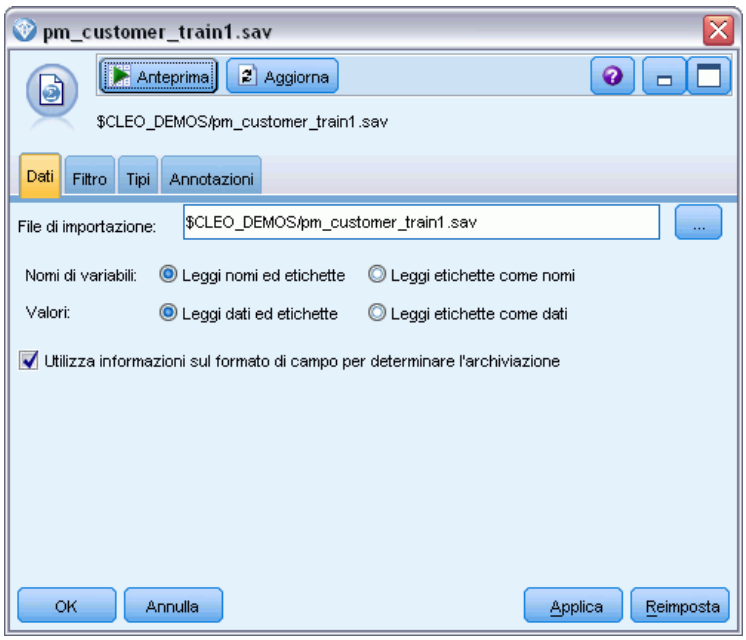
I valori del campo *campagna* vengono codificati come numeri interi nei dati, con etichette definite nel nodo Tipo (per esempio, *2 = Premium account*). È possibile attivare/disattivare la visualizzazione delle etichette dei valori nella tabella utilizzando la barra degli strumenti.

Il file include inoltre un numero di campi contenenti informazioni finanziarie e demografiche relative a ciascun cliente che possono essere utilizzate per generare o "addestrare" un modello che prevede i tassi di risposta relativi a gruppi di clienti diversi in base a caratteristiche specifiche.

Creazione dello stream

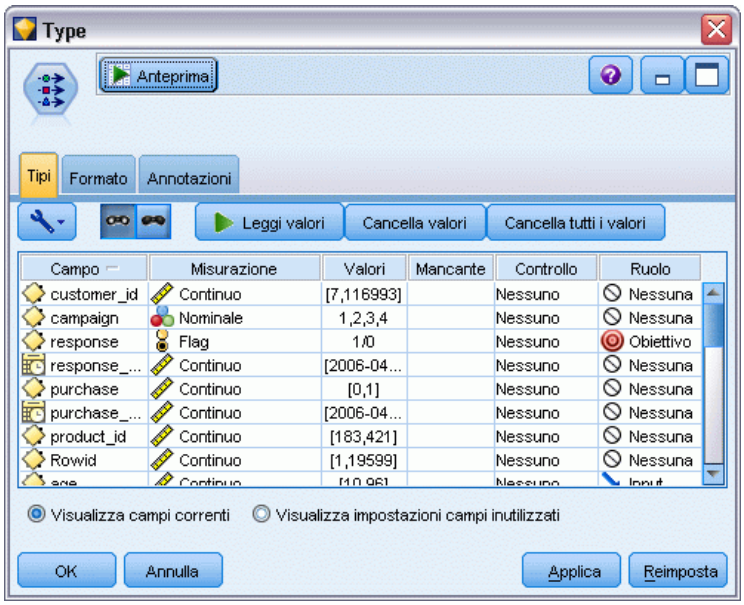
- Aggiungere un nodo File Statistics che punti al file *pm_customer_train1.sav* situato nella directory *Demos* dell'installazione IBM® SPSS® Modeler. In alternativa, per fare riferimento a questa cartella, è possibile specificare $CLEO_DEMOS/ nel percorso di file.

Figura 11-3
Lettura dei dati



- Aggiungere un nodo Tipo e selezionare *risposta* come campo obiettivo (Ruolo = Obiettivo). Impostare il livello di misurazione di questo campo su Flag.

Figura 11-4
Impostazione del livello di misurazione e del ruolo

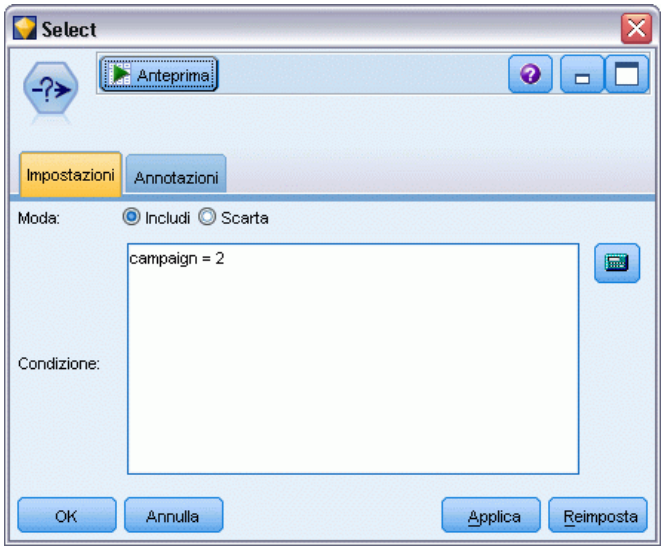


- Impostare su Nessuno il ruolo dei seguenti campi: *id_cliente*, *campagna*, *data_risposta*, *acquisto*, *data_acquisto*, *id_prodotto*, *IDriga* e *X_casuale*. Tutti questi campi hanno una propria funzione nei dati, tuttavia non vengono utilizzati per generare il modello.

- Fare clic sul pulsante Leggi valori nel nodo Tipo per assicurarsi che tali valori vengano istanziati.

Sebbene i dati includano informazioni su quattro diverse campagne, l'analisi verrà focalizzata su una campagna alla volta. Poiché il numero maggiore di record rientra nella campagna Premium account (codificata come *campagna = 2* nei dati), è possibile utilizzare il nodo Seleziona per includere nello stream solo i suddetti record.

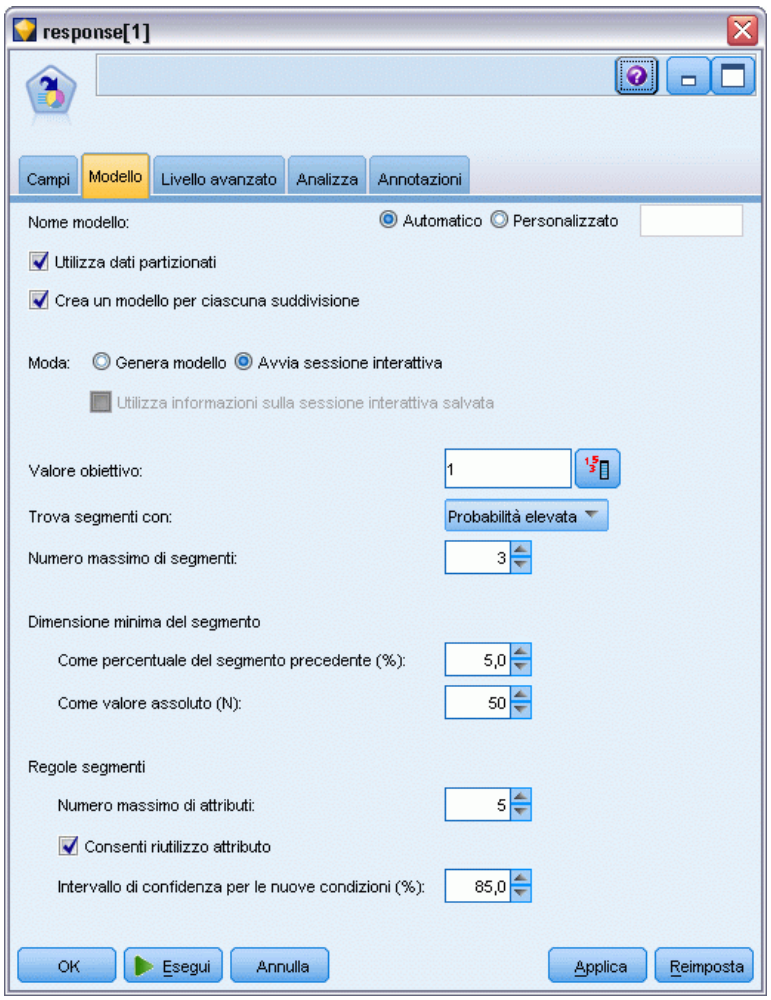
Figura 11-5
Selezione di record per un'unica campagna



Creazione del modello

- Collegare un nodo Elenco decisionale allo stream. Nella scheda Modello, impostare il valore obiettivo su 1, per indicare il risultato che si desidera cercare. In questo caso, si stanno cercando i clienti che hanno risposto *Sì* a un'offerta precedente.

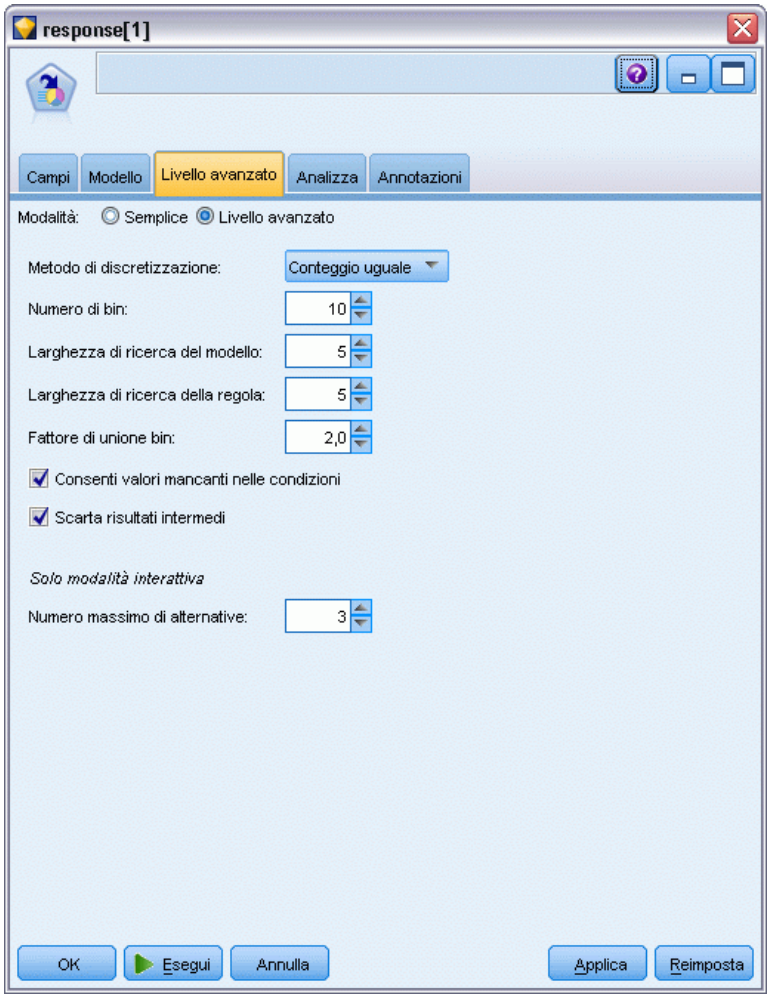
Figura 11-6
Scheda Modello del nodo Elenco decisionale



- Selezionare Avvia sessione interattiva.
- Per semplificare il modello ai fini del presente esempio, impostare il numero massimo di segmenti su 3.
- Modificare l'intervallo di confidenza per le nuove condizioni su 85%.

▶ Nella scheda Livello avanzato, impostare la Modalità su Livello avanzato.

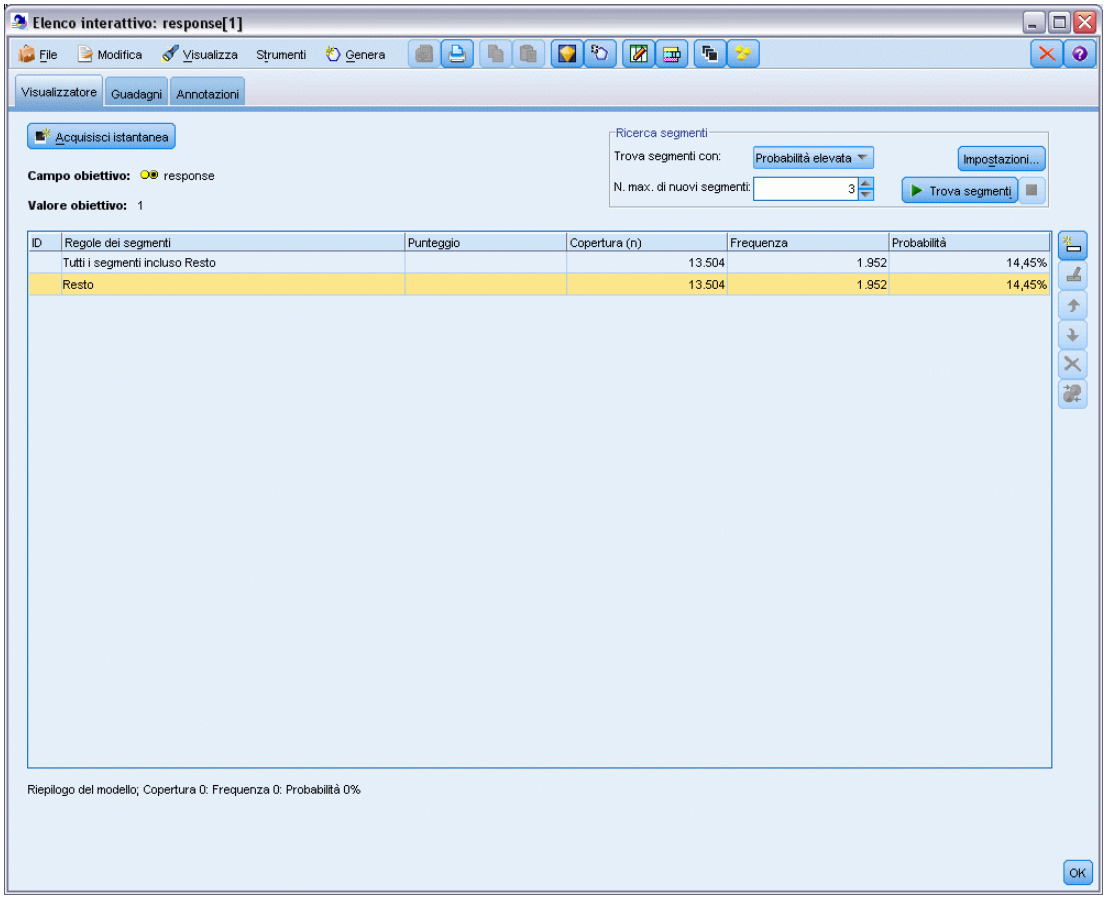
Figura 11-7
Nodo Elenco decisionale, scheda Livello avanzato



▶ Aumentare il Numero massimo di alternative a 3. Questa opzione funziona in combinazione con l'impostazione Avvia sessione interattiva selezionata nella scheda Modelli.

▶ Fare clic su Esegui per aprire il visualizzatore Elenco interattivo.

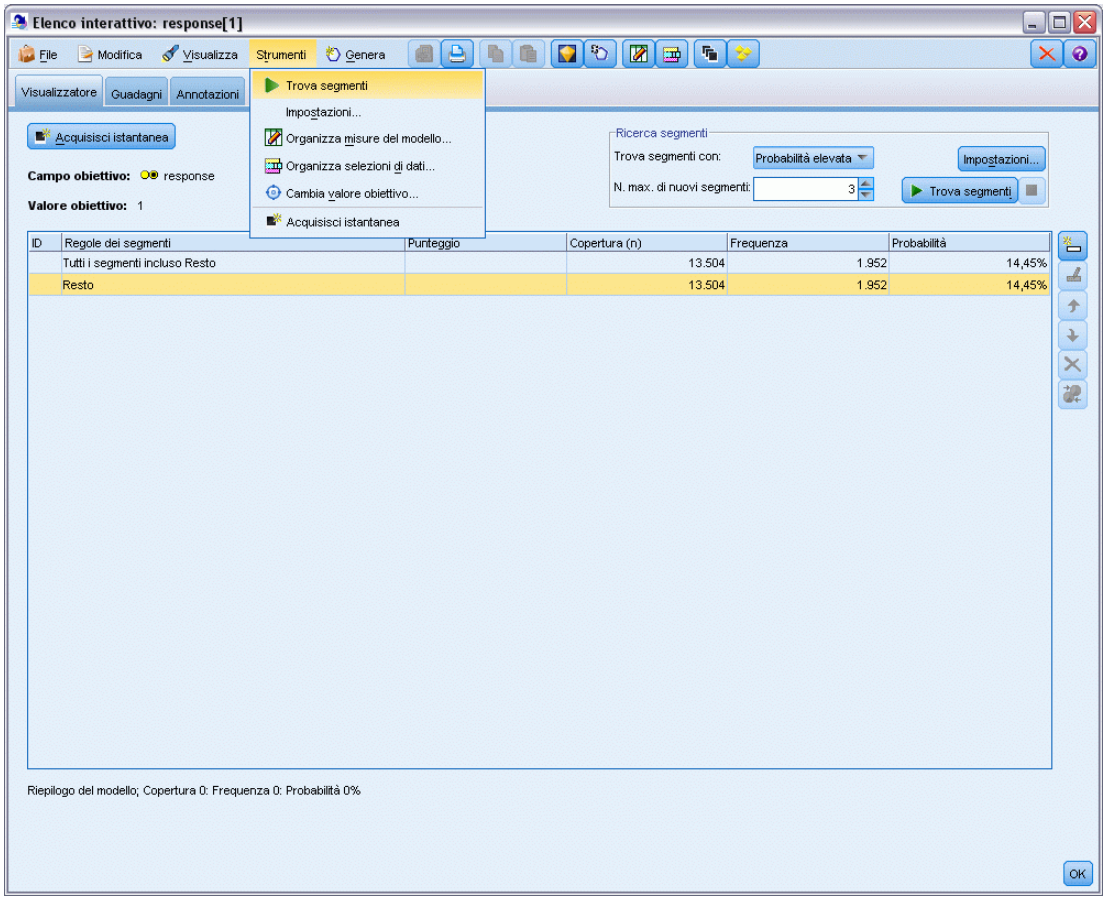
Figura 11-8
visualizzatore Elenco interattivo



Poiché non è stato ancora definito alcun segmento, tutti i record rientrano nella categoria resto. Dei 13.504 record presenti nel campione, 1.952 hanno risposto *Sì*, per un tasso di risultati positivi globale pari a 14,45%. Per migliorare questo tasso, è possibile identificare i segmenti di clienti che hanno più (o meno) probabilità di rispondere in modo positivo.

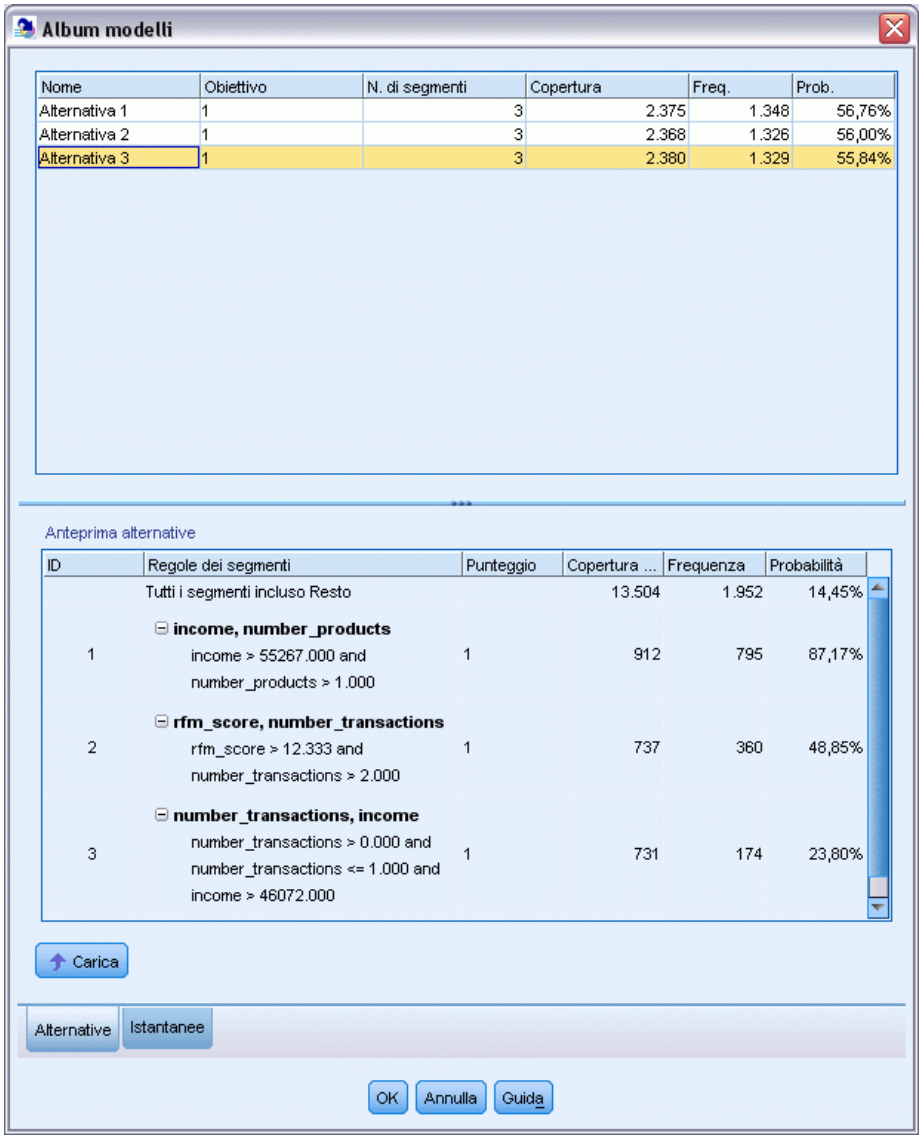
▶ Dai menu del visualizzatore Elenco interattivo, scegliere:
Strumenti > Trova segmenti

Figura 11-9
visualizzatore Elenco interattivo



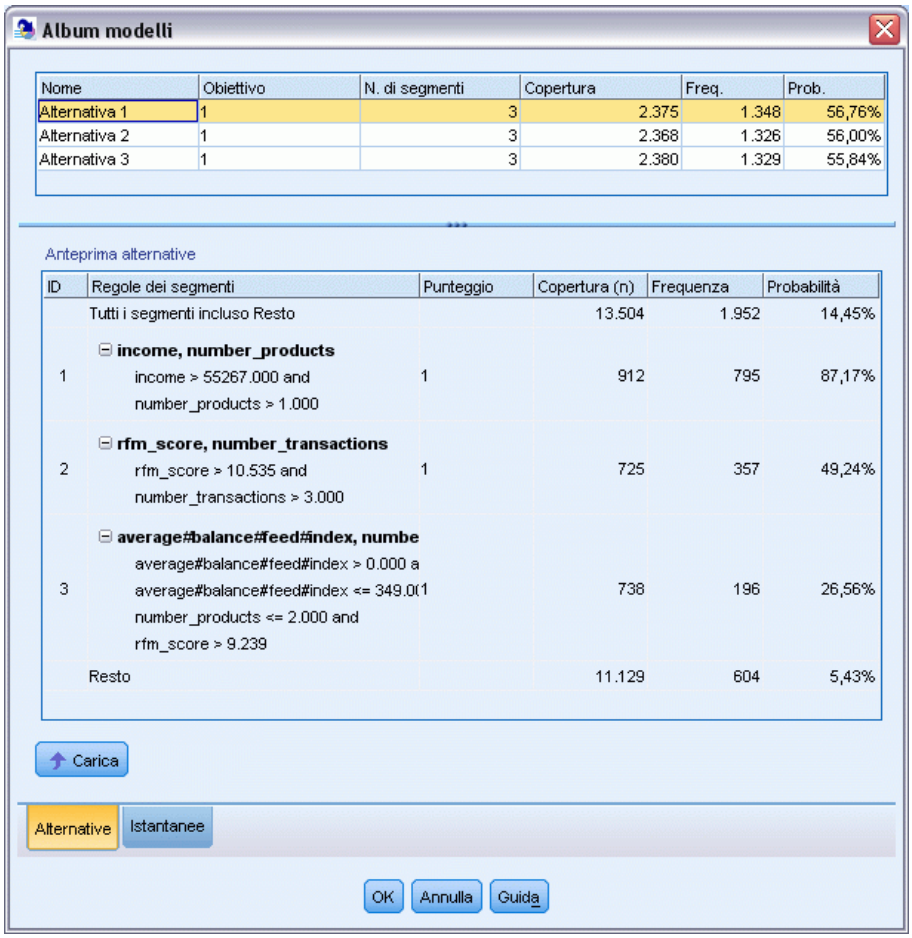
In questo modo, viene eseguita l'attività di mining di default in base alle impostazioni specificate nel nodo Elenco decisionale. L'attività completata restituisce tre modelli alternativi, che vengono elencati nella scheda Alternative della finestra di dialogo Album modelli.

Figura 11-10
Modelli alternativi disponibili



- Selezionare la prima alternativa dall'elenco; i relativi dettagli sono visualizzati nel riquadro Anteprima alternative.

Figura 11-11
Modello alternativo selezionato



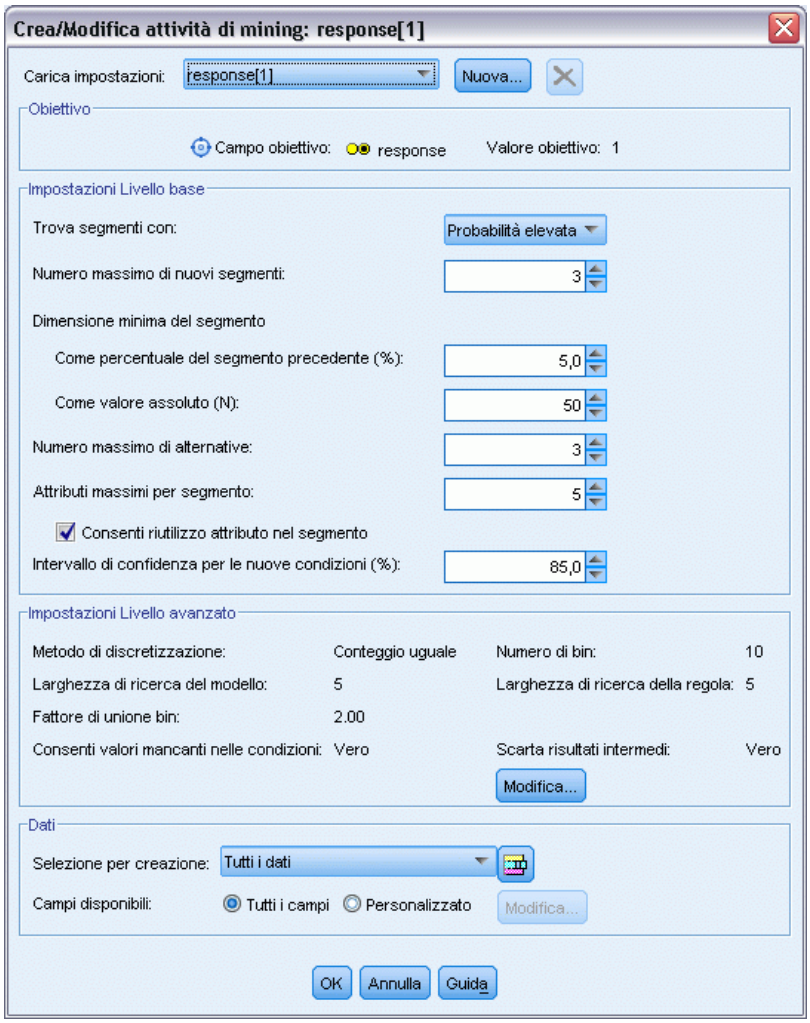
Il riquadro Anteprima alternative consente di scorrere rapidamente numerose alternative senza cambiare modello di lavoro, permettendo di sperimentare facilmente vari approcci.

Nota: per visualizzare meglio il modello, è possibile ingrandire il riquadro Anteprima alternative all'interno della finestra, come indicato di seguito. È possibile effettuare questa operazione trascinando il bordo del riquadro.

Utilizzando regole basate su predittori quali reddito, numero delle transazioni mensili e punteggio RFM, il modello identifica i segmenti con tassi di risposta più alti rispetto a quelli del campione nel suo insieme. Quando i segmenti vengono combinati, questo modello suggerisce che è possibile migliorare il tasso di risposte positive a 56,76%. Tuttavia, il modello copre solo una piccola parte dell'intero campione e quasi 11.000 record, comprese diverse centinaia di risultati positivi, rimangono nella categoria resto. È necessario quindi utilizzare un modello che catturi un numero maggiore di risultati positivi, pur continuando a escludere i segmenti a performance ridotta.

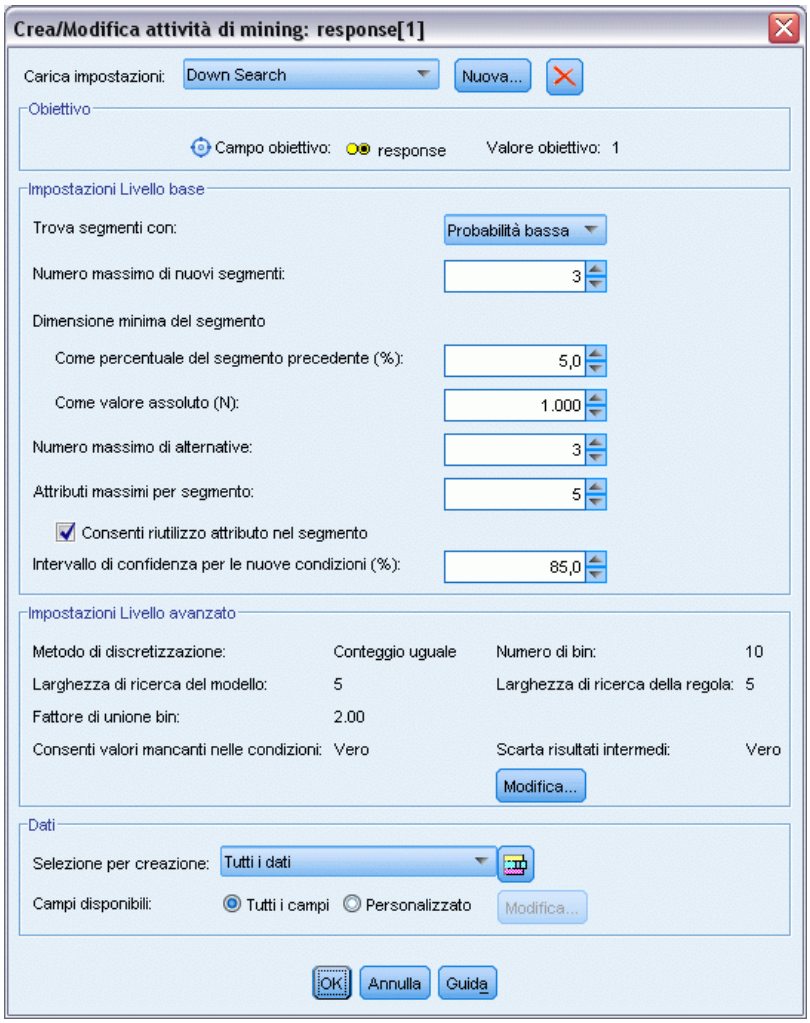
▶ Per provare un approccio di modellazione diverso, effettuare le seguenti scelte dai menu:
Strumenti > Impostazioni

Figura 11-12
Finestra di dialogo Crea/Modifica attività di mining



▶ Fare clic sul pulsante Nuovo (angolo superiore destro) per creare una seconda attività di mining, quindi specificare *Cerca in basso* come nome dell'attività nella finestra di dialogo Nuove impostazioni.

Figura 11-13
Finestra di dialogo Crea/Modifica attività di mining

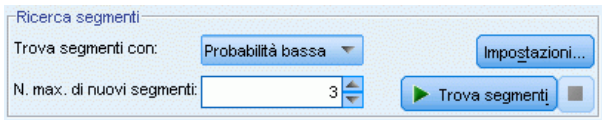


- Cambiare la direzione di ricerca su Probabilità bassa per l'attività. In questo modo, l'algoritmo cercherà i segmenti con i tassi di risposta *più bassi* invece che quelli con i tassi più alti.

- Aumentare la dimensione minima dei segmenti a 1.000. Fare clic su OK per tornare al visualizzatore Elenco interattivo.

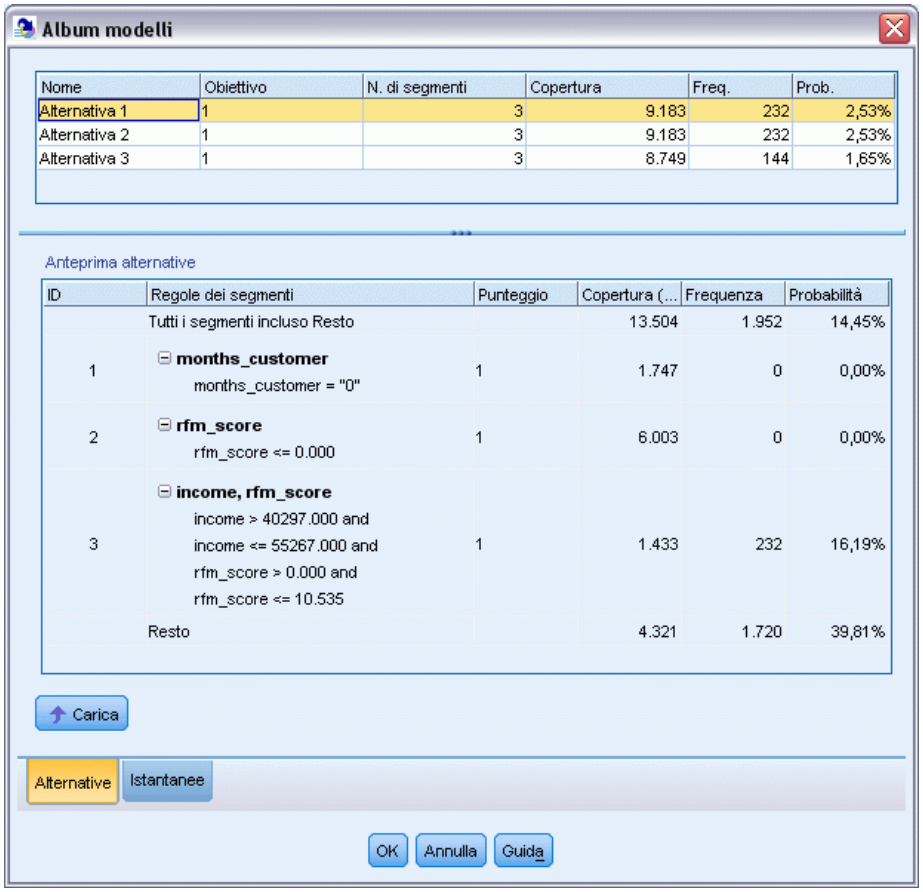
- Nel visualizzatore Elenco interattivo, verificare che il riquadro *Ricerca segmenti* visualizzi i dettagli della nuova attività e fare clic su Trova segmenti.

Figura 11-14
Ricerca di segmenti in una nuova attività di mining



L'attività restituisce una nuova serie di alternative, che vengono visualizzate nella scheda Alternative della finestra Album modelli e che possono essere visualizzate in anteprima analogamente ai risultati precedenti.

Figura 11-15
Risultati del modello Cerca in basso

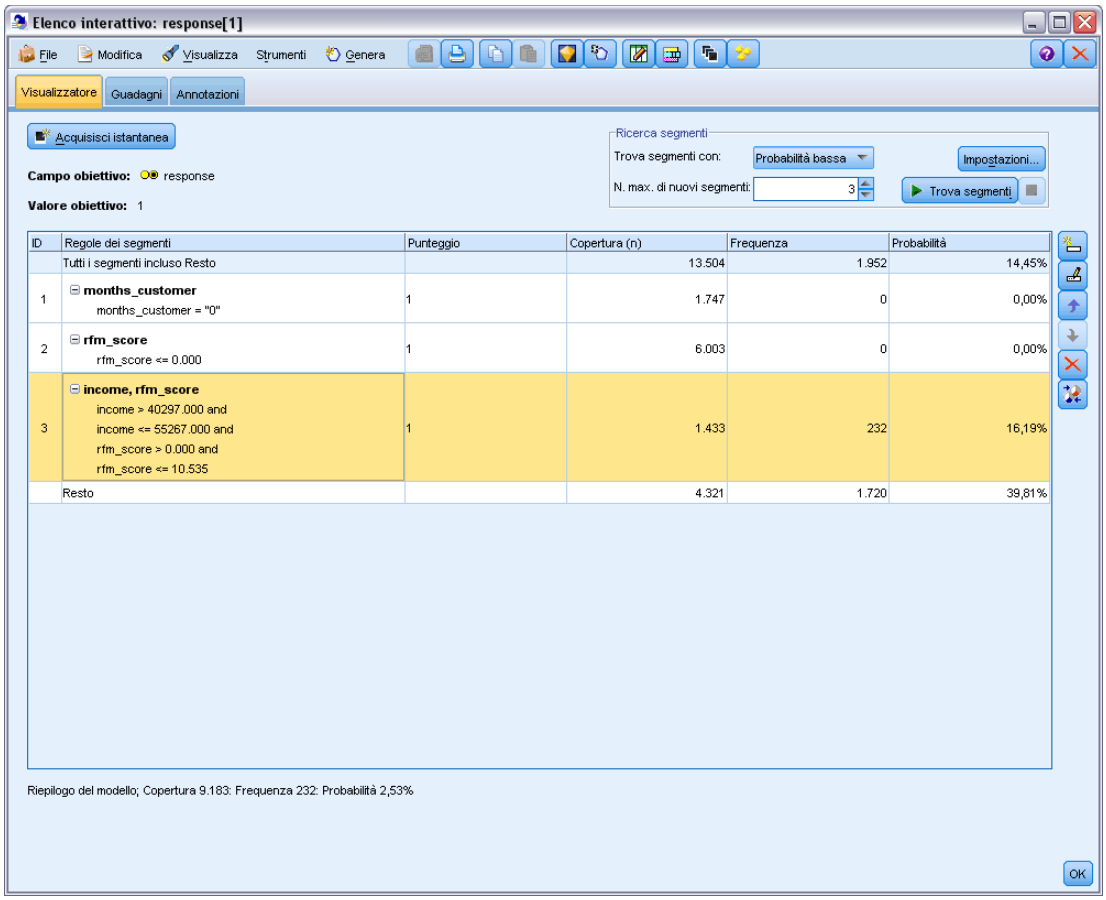


Questa volta, ogni modello identifica i segmenti con le probabilità di risposta più basse invece che più alte. Osservando la prima alternativa, la semplice esclusione di questi segmenti consente di incrementare il tasso di risultati positivi del resto al 39,81%. Questo risultato è inferiore rispetto al modello osservato in precedenza, ma implica una copertura più alta (un numero totale di risultati positivi più alto).

Combinando i due approcci, ovvero utilizzando una ricerca Probabilità bassa per eliminare i record non significativi seguita da una ricerca Probabilità elevata, è possibile migliorare questo risultato.

- Fare clic su Carica per rendere questo modello (la prima alternativa di Cerca in basso) il modello di lavoro e fare clic su OK per chiudere l'Album modelli.

Figura 11-16
Esclusione di un segmento



- Fare clic con il pulsante destro del mouse su ciascuno dei primi due segmenti e scegliere Escludi segmento. Entrambi questi segmenti catturano circa 8.000 record con risultati positivi zero, quindi sembra più che appropriato escluderli dalle offerte future. Ai segmenti esclusi verrà assegnato un punteggio nullo, in modo da indicare l'esclusione.

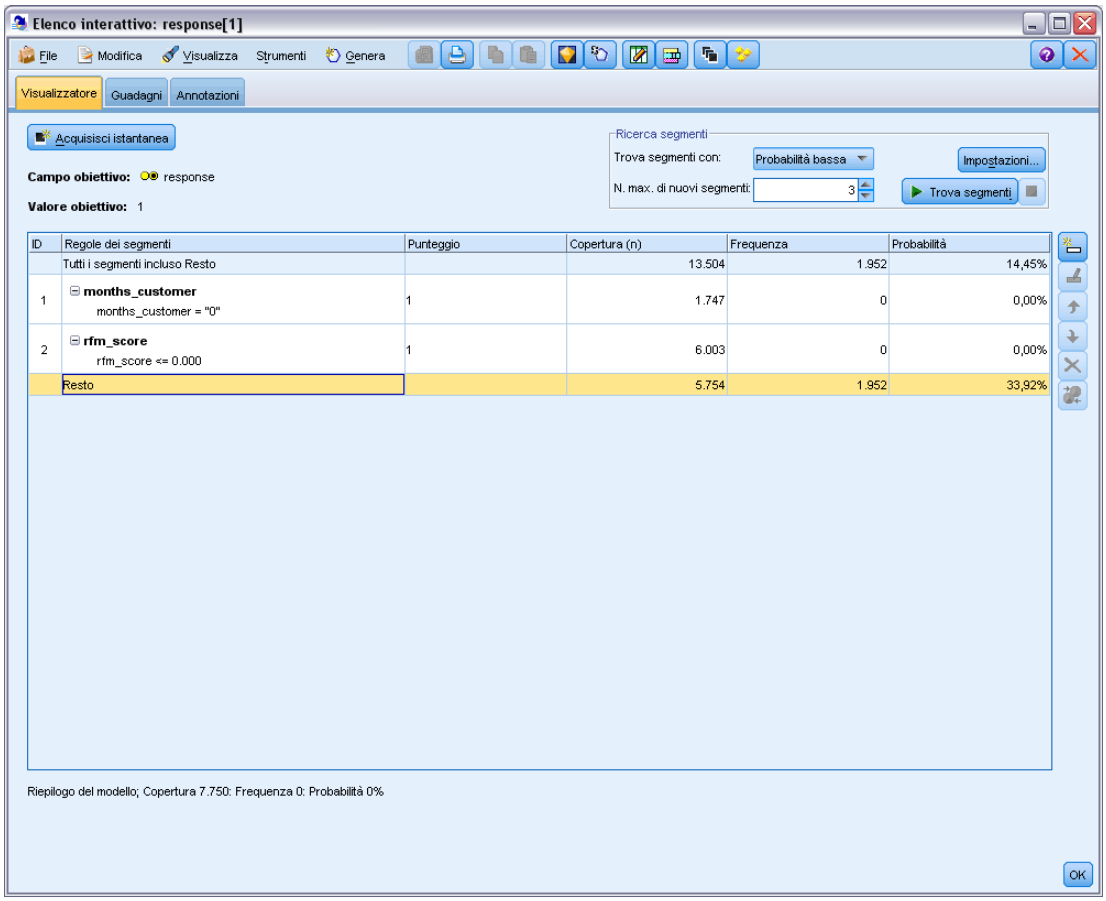
- Fare clic con il pulsante destro del mouse sul terzo segmento e selezionare Elimina segmento. Il tasso di risultati positivi del 16,19% di questo segmento non è molto diverso dal tasso di riferimento del 14,45%, pertanto il segmento non aggiunge informazioni sufficienti da giustificarne il mantenimento.

  Nota: l'eliminazione di un segmento è diversa dall'esclusione. L'esclusione di un segmento modifica semplicemente il metodo di calcolo del relativo punteggio, mentre l'eliminazione comporta la rimozione del segmento dal modello.

  Dopo aver escluso i segmenti con le performance più basse, si cercheranno fra i rimanenti i segmenti che presentano le performance più alte.

- Fare clic sulla riga del resto nella tabella per selezionarla, in modo che la prossima esecuzione dell'attività di mining venga applicata solo al resto.

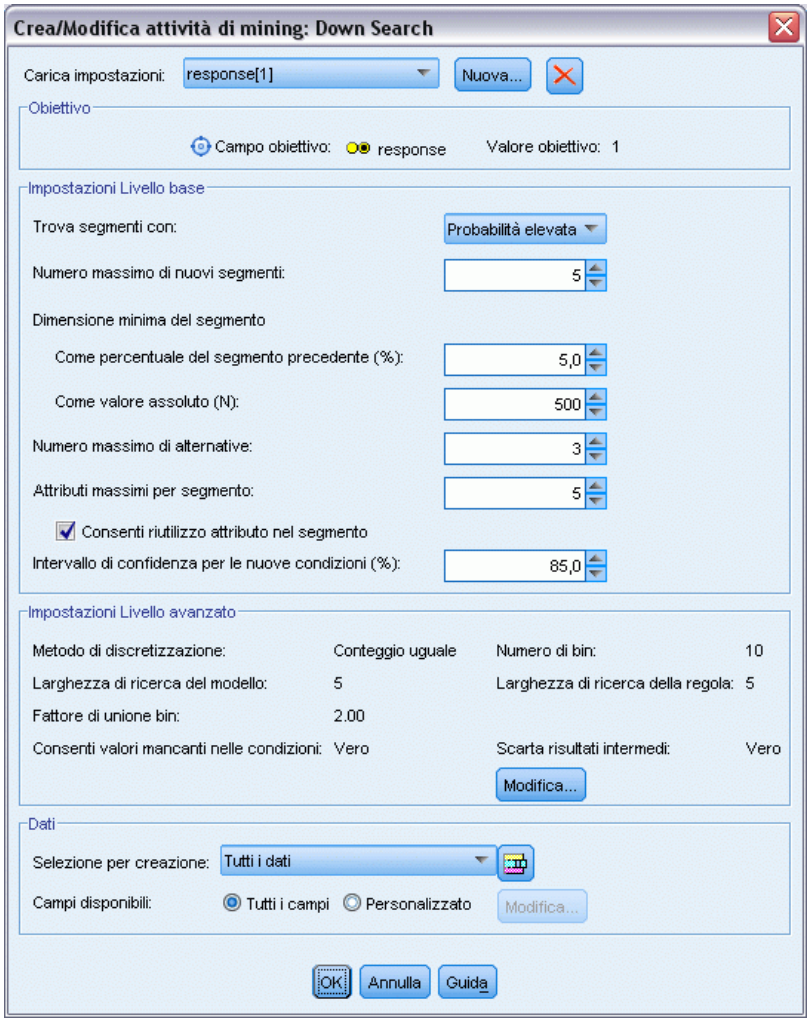
Figura 11-17
Selezione di un segmento



- Dopo aver selezionato la riga del resto, fare clic su Impostazioni per riaprire la finestra di dialogo Crea/Modifica attività di mining.
- Nella parte superiore, in Carica impostazioni, selezionare l'attività di mining di default: risposta[1].
- Modificare le Impostazioni Livello base per portare a 5 il numero dei nuovi segmenti e a 500 la dimensione minima dei segmenti.

▶ Fare clic su OK per tornare al visualizzatore Elenco interattivo.

Figura 11-18
Selezione dell'attività di mining di default

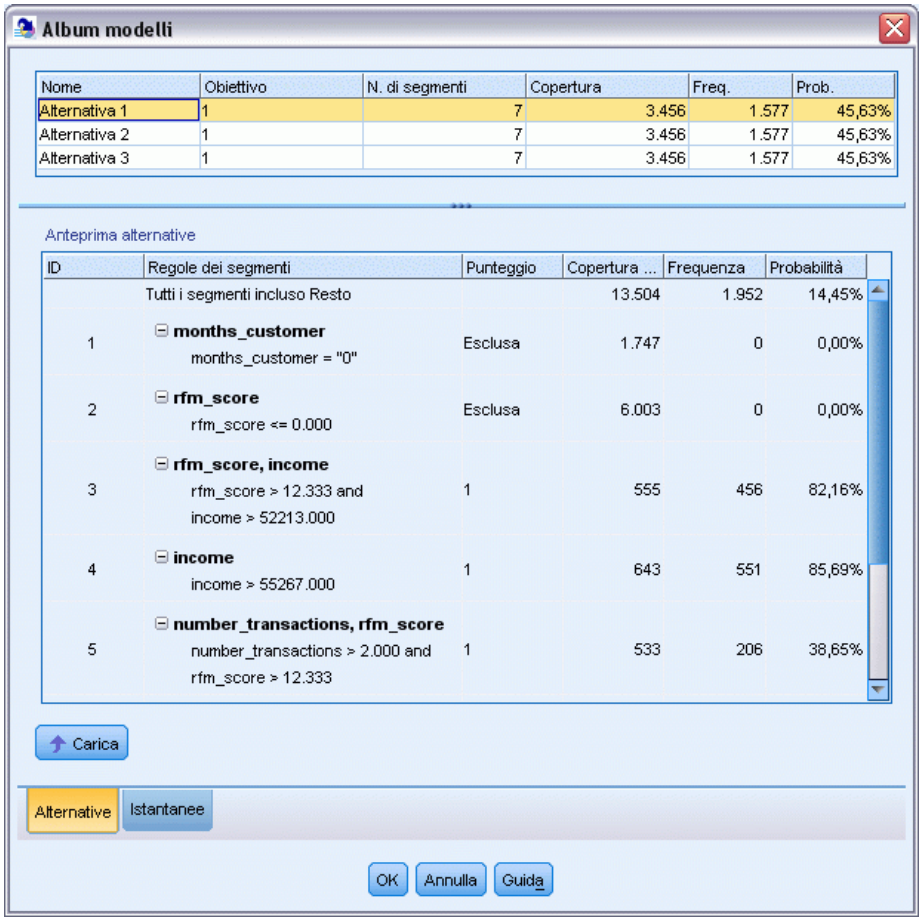


▶ Fare clic su Trova segmenti.

In questo modo viene visualizzata un'altra serie di modelli alternativi. Poiché i risultati di un'attività di mining sono stati introdotti in un'altra attività, questi ultimi modelli contengono una combinazione di segmenti ad alte e basse performance. I segmenti con bassi tassi di risposta vengono esclusi, il che significa che verranno loro assegnati dei punteggi nulli, mentre ai segmenti inclusi verrà assegnato il punteggio 1. Le statistiche globali riflettono queste esclusioni, con il

modello della prima alternativa che evidenzia un tasso di risultati positivi pari a 45,63%, con una copertura più alta (1.577 risultati positivi su 3.456 record) rispetto a tutti i modelli precedenti.

Figura 11-19
Alternative per modello combinato

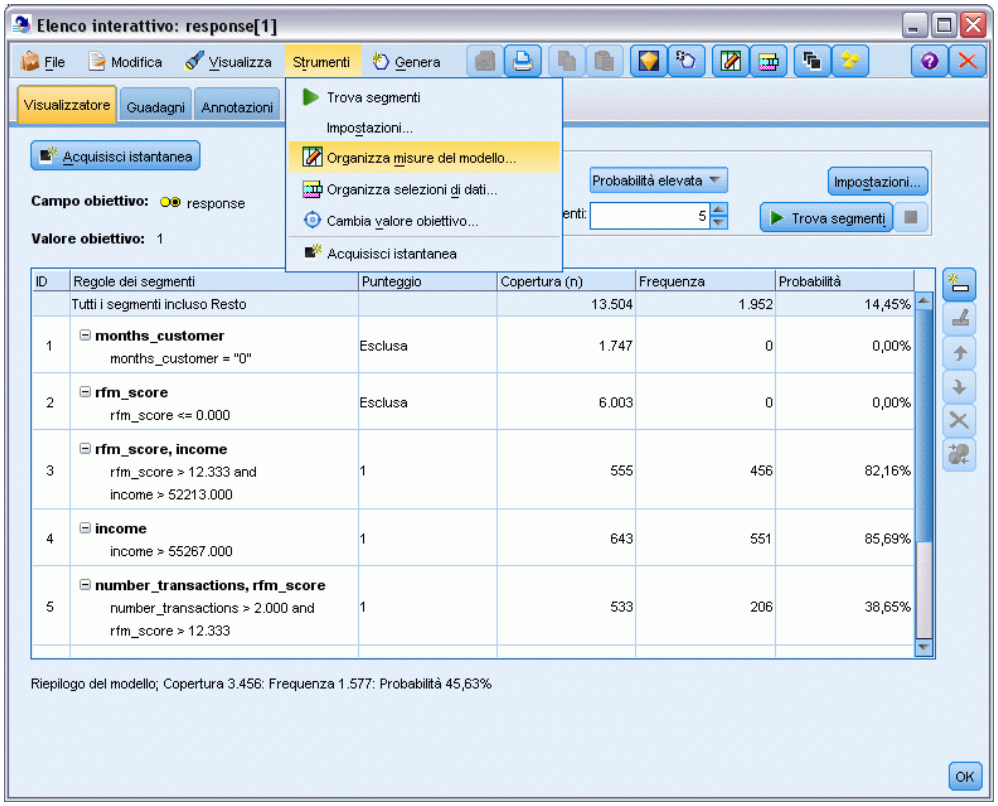


- Visualizzare in anteprima la prima alternativa e fare clic su Carica per rendere questo modello il modello di lavoro.

Calcolo di misure personalizzate con Excel

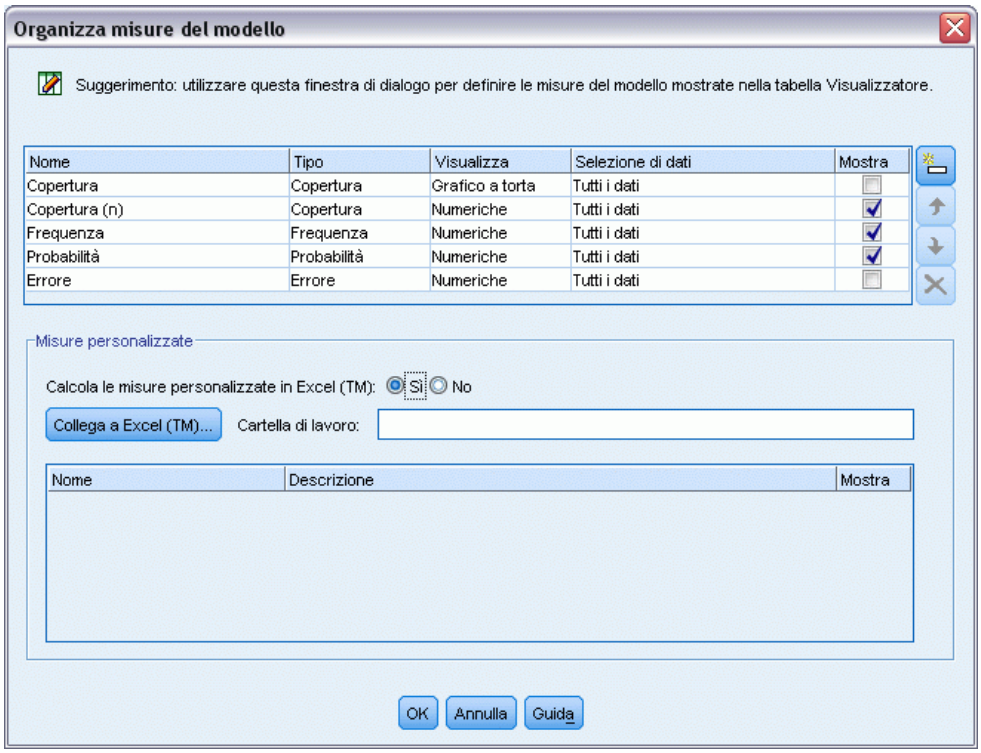
- Per acquisire maggiore familiarità con il funzionamento del modello in termini pratici, scegliere Organizza misure del modello dal menu Strumenti.

Figura 11-20
Organizzazione delle misure del modello



La finestra di dialogo Organizza misure del modello consente di scegliere le misure (o colonne) da visualizzare nel visualizzatore Elenco interattivo. È inoltre possibile specificare quali misure vengono calcolate rispetto a tutti i record o a un sottoinsieme selezionato e scegliere di visualizzare un grafico a torta anziché un numero, laddove possibile.

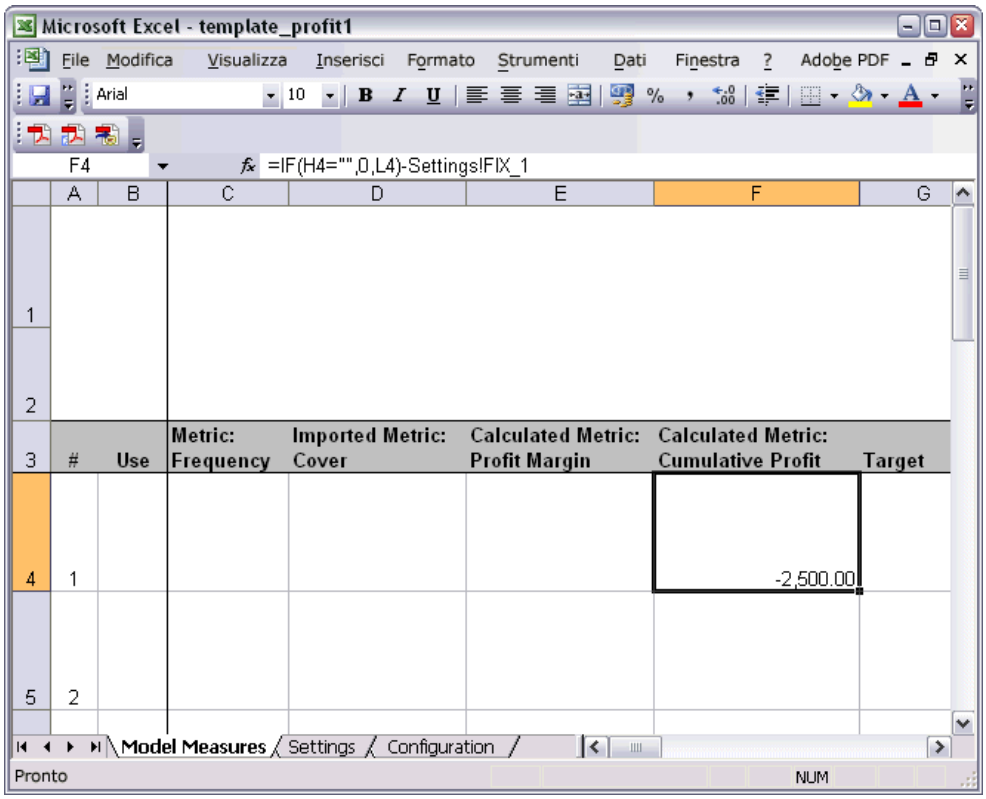
Figura 11-21
Finestra di dialogo Organizza misure del modello



Inoltre, se sul computer è installato Microsoft Excel, è possibile collegarsi a un modello di Excel che calcolerà le misure personalizzate e le aggiungerà alla visualizzazione interattiva.

- Nella finestra di dialogo Organizza misure del modello, impostare Calcola le misure personalizzate in Excel (TM) su Sì.
- Fare clic su Collega a Excel (TM).
- Selezionare la cartella di lavoro *template_profit.xlt* che si trova in *streams* nella cartella *Demos* dell'installazione IBM® SPSS® Modeler, quindi scegliere Apri per avviare il foglio di calcolo.

Figura 11-22
Foglio di lavoro Excel delle misure del modello



Il modello di Excel contiene tre fogli di lavoro:

- Misure del modello visualizza le misure del modello importate dal modello e calcola misure personalizzate per la successiva riesportazione nel modello.
- Impostazioni contiene parametri da utilizzare per il calcolo di misure personalizzate.
- Configurazione definisce le misure da importare da ed esportare nel modello.

Di seguito sono elencate le metriche riesportate nel modello:

- **Margine di profitto.** Entrata netta dal segmento
- **Profitto cumulato.** Profitto totale generato dalla campagna

Come definito dalle seguenti formule:

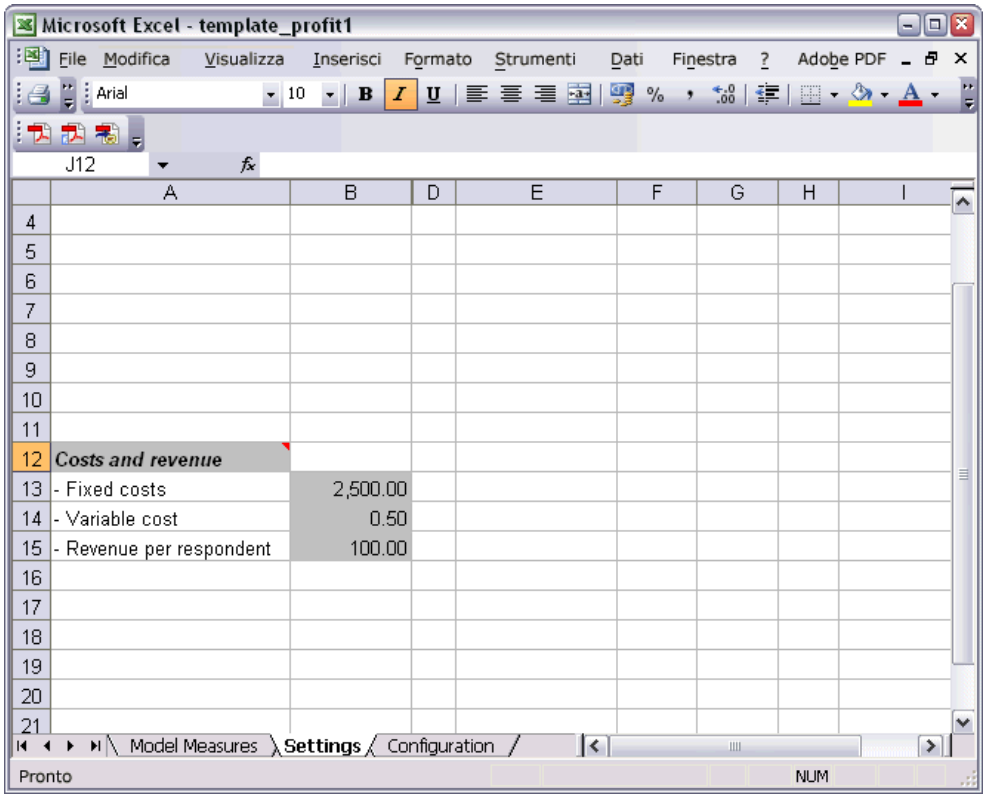
Profit Margin = Frequency * Revenue per respondent - Cover * Variable cost

Cumulative Profit = Total Profit Margin - Fixed cost

Si noti che Frequenza e Copertura vengono importati dal modello.

I parametri Costo ed Entrate vengono specificati dall'utente nel foglio di lavoro Impostazioni.

Figura 11-23
Foglio di lavoro Impostazioni di Excel



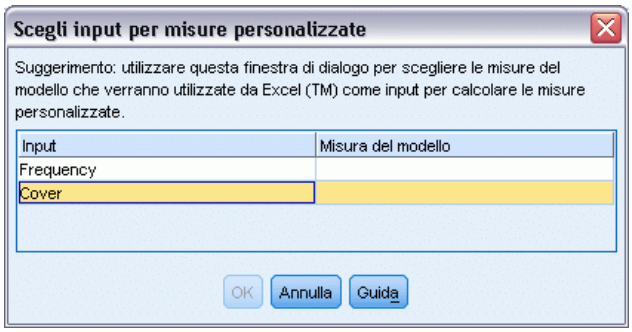
Il **Costo fisso** è il costo di allestimento della campagna, quale il costo di progettazione e pianificazione.

Il **Costo variabile** è il costo di estensione dell'offerta a ogni cliente; per esempio il costo di buste e francobolli.

L'**Entrata per rispondente** è l'entrata netta da un cliente che risponde all'offerta.

- Per concludere il collegamento al modello, utilizzare la barra delle applicazioni di Windows (oppure premere Alt+Tab) per ritornare al visualizzatore Elenco interattivo.

Figura 11-24
Scelta di input per misure personalizzate



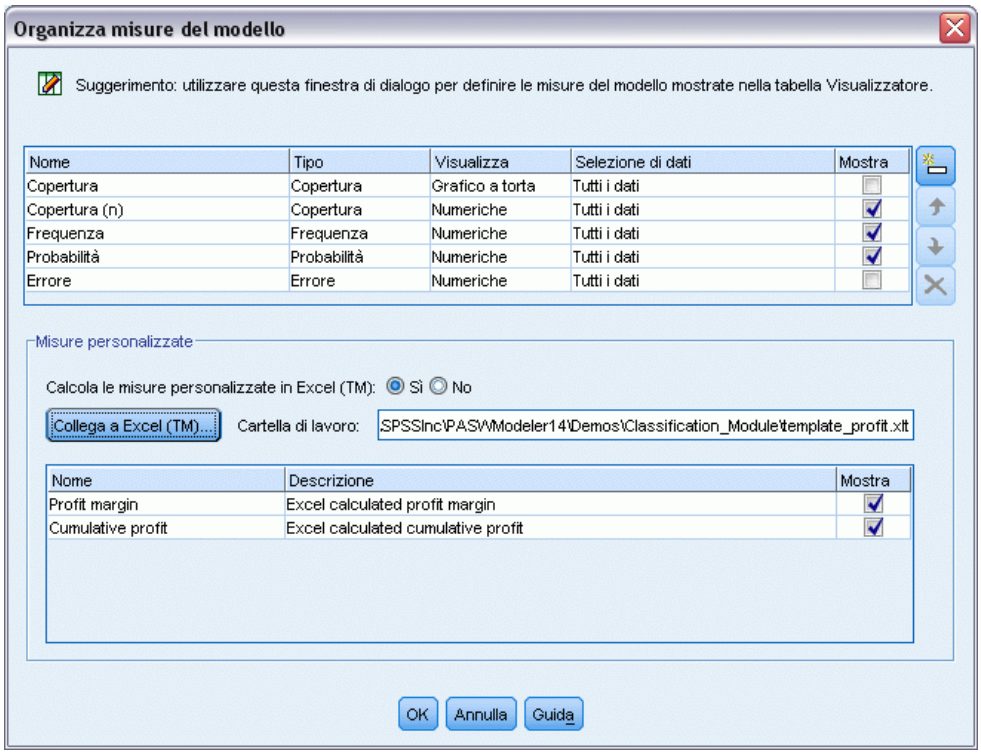
Viene visualizzata la finestra di dialogo Scegli input per misure personalizzate che consente di associare input dal modello a parametri specifici definiti nel modello di Excel. Nella colonna di sinistra sono elencate tutte le misure disponibili, mentre in quella di destra tali misure sono associate a parametri del foglio di lavoro, come definito nel foglio di lavoro Configurazione.

- Nella colonna Misure del modello, selezionare Frequenza e Copertura (n) relativamente ai rispettivi input, quindi fare clic su OK.

In questo caso, è capitato che i nomi dei parametri nel modello, Frequenza e Copertura (n), corrispondessero agli input, ma potrebbero venire utilizzati nomi diversi.

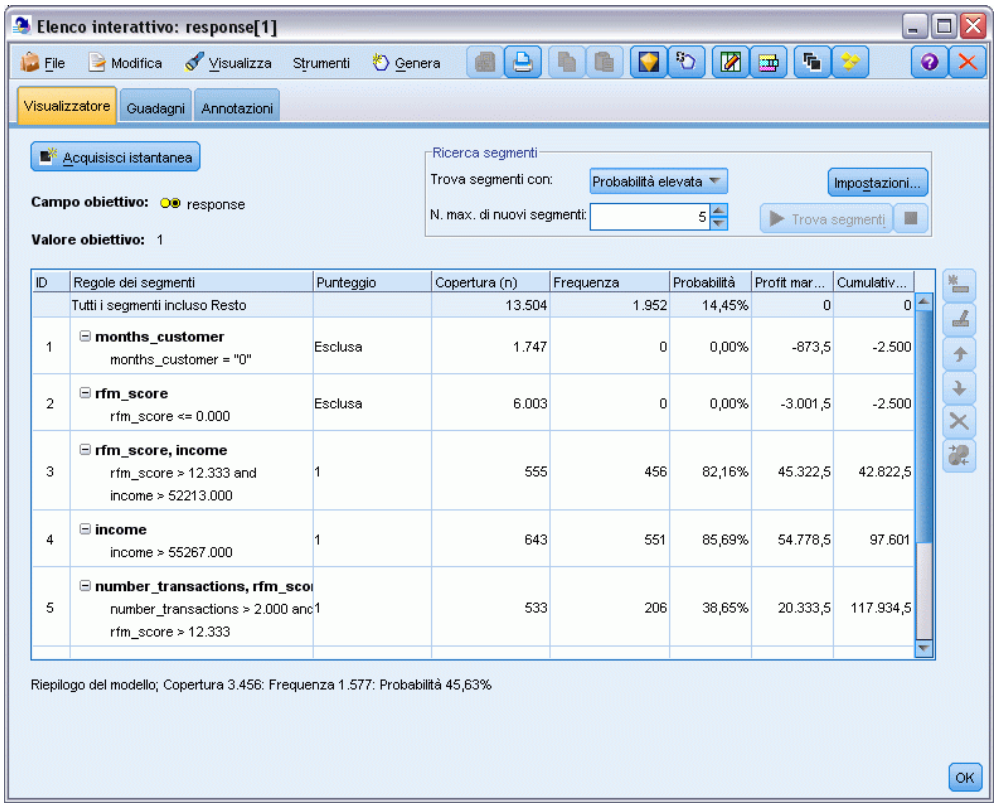
▶ Fare clic su OK nella finestra di dialogo Organizza misure del modello per aggiornare il visualizzatore Elenco interattivo.

Figura 11-25
Finestra di dialogo Organizza misure del modello che mostra le misure personalizzate da Excel



Le nuove misure vengono ora aggiunte alla finestra come nuove colonne e vengono ricalcolate a ogni nuovo aggiornamento del modello.

Figura 11-26
Misure personalizzate da Excel visualizzate nel visualizzatore Elenco interattivo



Modificando il modello di Excel, è possibile creare un numero qualsiasi di misure personalizzate.

Modifica del modello Excel

Sebbene IBM® SPSS® Modeler venga fornito con un modello Excel di default da utilizzare con il visualizzatore Elenco interattivo, è possibile modificarne le impostazioni o aggiungerne di proprie. Per esempio, i costi inclusi nel modello potrebbero non essere corretti per l'organizzazione e potrebbe essere pertanto necessario modificarli.

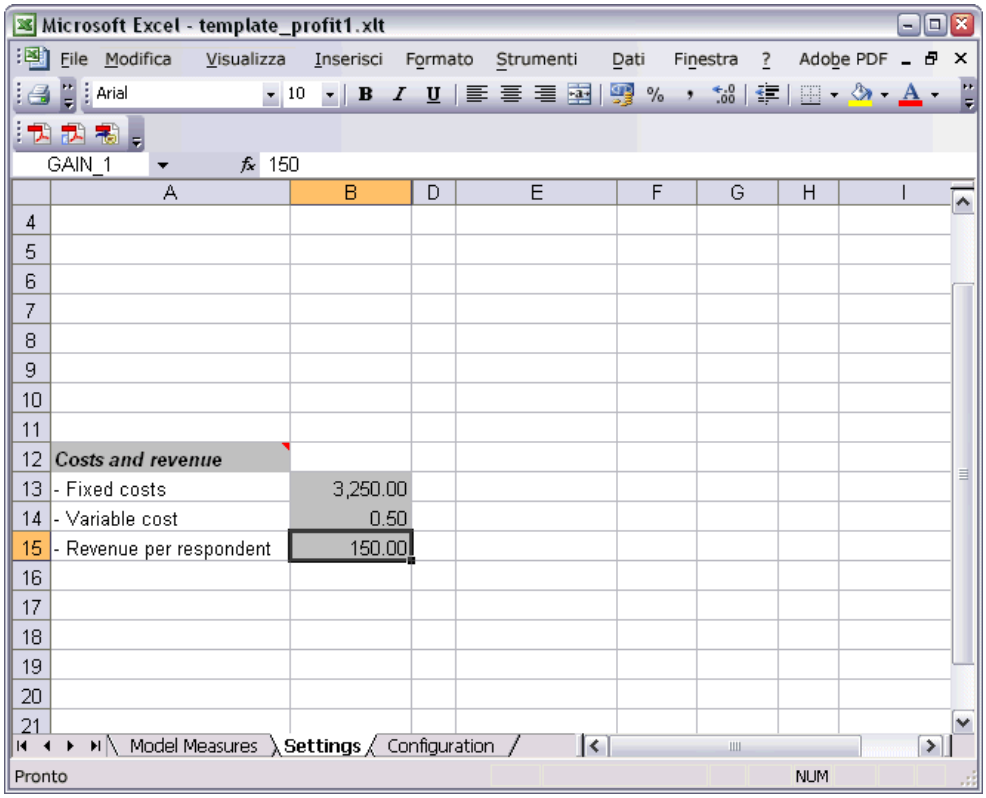
Nota: se si modifica un modello esistente o se ne crea uno proprio, è necessario salvare il file con il suffisso *.xlt* di Excel 2003.

Per modificare il modello predefinito con nuovi dettagli di costi ed entrate e aggiornare il visualizzatore Elenco interattivo con le nuove cifre:

- Nel visualizzatore Elenco interattivo, scegliere Organizza misure del modello dal menu Strumenti.
- Nella finestra di dialogo Organizza misure del modello, fare clic su Collega a Excel™.
- Selezionare la cartella di lavoro *template_profit.xlt* e fare clic su Apri per avviare il foglio di calcolo.

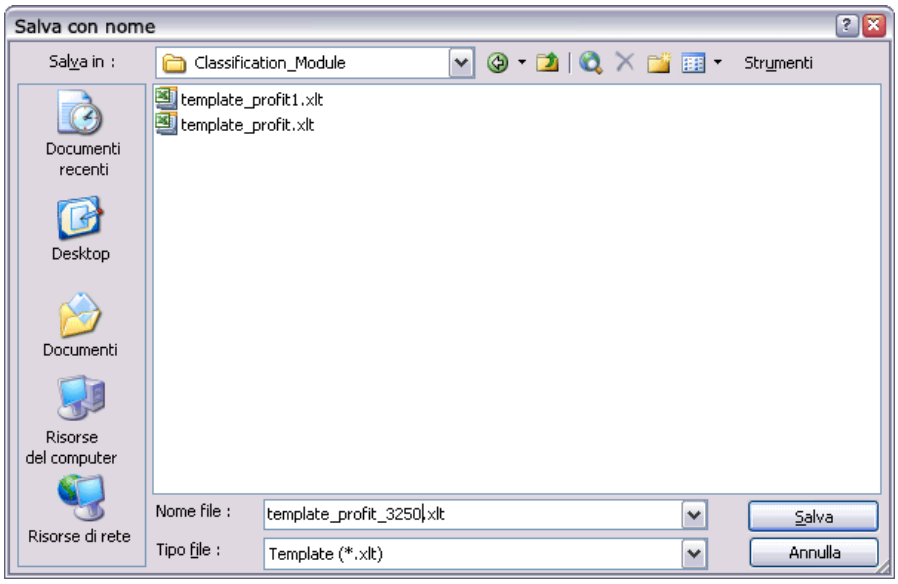
- Selezionare il foglio di lavoro Impostazioni.
- Modificare il Costo fisso impostandolo su 3.250,00 e l'Entrata per rispondente impostandola su 150,00.

Figura 11-27
Modifica di valori nel foglio di calcolo Impostazioni di Excel



- Salvare il modello modificato con un nome di file pertinente e univoco. Assicurarsi che abbia l'estensione *.xlt* di Excel 2003.

Figura 11-28
Salvataggio del modello Excel modificato



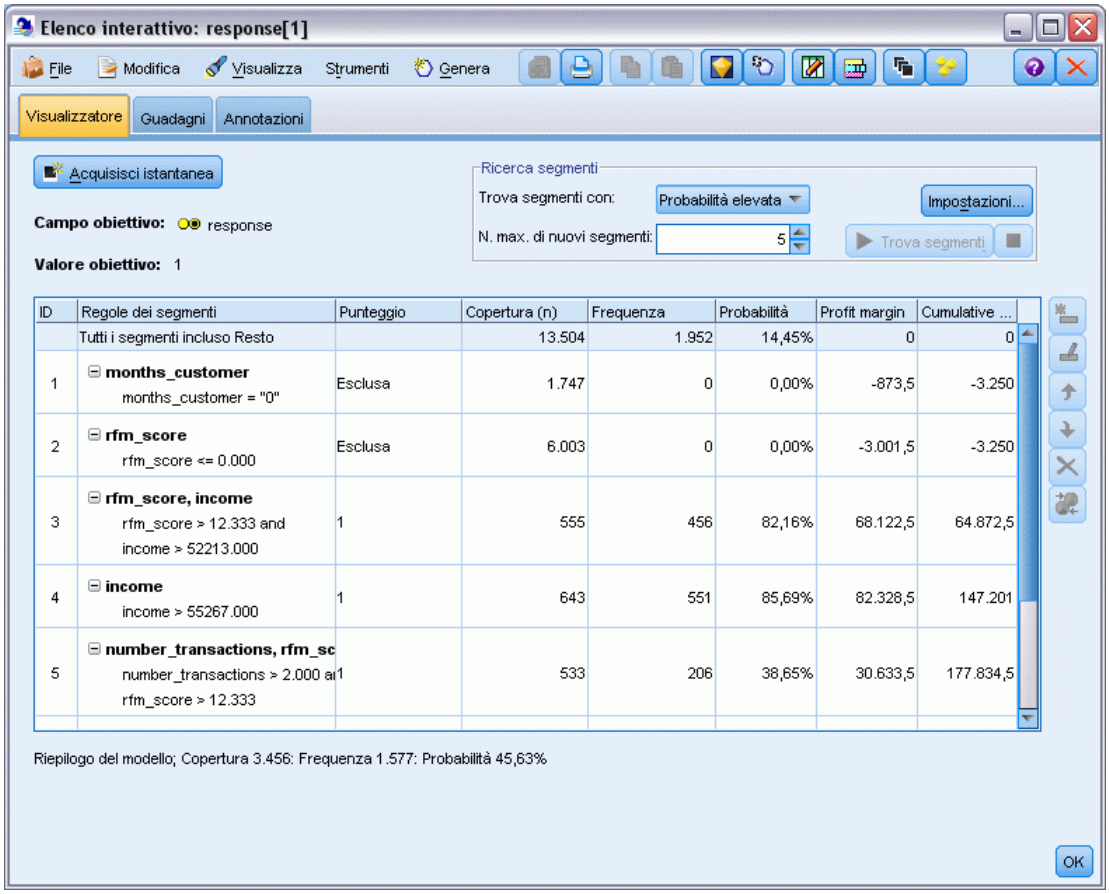
- Utilizzare la barra delle applicazioni di Windows (oppure premere Alt+Tab) per ritornare al visualizzatore Elenco interattivo.

  Nella finestra di dialogo Scegli input per misure personalizzate, selezionare le misure da visualizzare e fare clic su OK.

- Nella finestra di dialogo Organizza misure del modello, fare clic su OK per aggiornare il visualizzatore Elenco interattivo.

Ovviamente, in questo esempio è stato illustrato solo un semplice modo per modificare il modello Excel. È possibile apportare ulteriori modifiche al modello per estrarre dati da e passare dati al visualizzatore Elenco interattivo o lavorare in Excel per produrre altri output, quali grafici.

Figura 11-29
Misure personalizzate modificate da Excel visualizzate nel visualizzatore Elenco interattivo



Salvataggio dei risultati

Per salvare un modello per l'utilizzo futuro durante la sessione interattiva, è possibile eseguire un'istantanea del modello, che verrà elencata nella scheda Istantanee. È possibile tornare alle istantanee salvate in qualsiasi momento durante la sessione interattiva.

Proseguendo in questo modo, è possibile sperimentare ulteriori attività di mining per cercare altri segmenti. Inoltre, è possibile modificare i segmenti esistenti, inserire segmenti personalizzati in base alle regole della propria azienda, creare selezioni di dati per ottimizzare il modello per gruppi specifici e personalizzare il modello in diversi altri modi. Infine, è possibile includere o escludere esplicitamente ciascun segmento secondo necessità e specificare come calcolarne il punteggio.

Dopo aver ottenuto risultati soddisfacenti, è possibile utilizzare il menu Genera per creare un modello che può essere aggiunto agli stream o sottoposto a deployment per il calcolo del punteggio.

In alternativa, se si desidera salvare lo stato attuale della propria sessione interattiva in modo da poterla riprendere in un secondo tempo, scegliere Aggiorna nodo Modelli dal menu File. In questo modo, il nodo Modelli Elenco decisionale verrà aggiornato con le impostazioni correnti, incluse le attività di mining, le istantanee dei modelli, le selezioni di dati e le misure personalizzate. Alla successiva esecuzione dello stream, sarà sufficiente verificare che sia stata selezionata l'opzione Utilizza informazioni sulla sessione interattiva salvata nel nodo Modelli Elenco decisionale per riportare la sessione allo stato attuale. Per ulteriori informazioni, vedere l'argomento Elenco decisionale in il capitolo 9 in *IBM SPSS Modeler 15 Nodi Modelli*.

Capitolo
12

Classificazione dei clienti nelle telecomunicazioni (Regressione logistica multinomiale)

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

Si supponga, per esempio, che una società che fornisce servizi di telecomunicazioni abbia segmentato la propria clientela in base a schemi di utilizzo del servizio, suddividendo i clienti in quattro gruppi. Se si utilizzano i dati demografici per prevedere l'appartenenza al gruppo, è possibile personalizzare le offerte per potenziali clienti individuali.

In questo esempio viene utilizzato lo stream denominato *telco_custcat.str*, che fa riferimento al file di dati denominato *telco.sav*. Questi file sono disponibili nella directory *Demos* di tutte le installazioni di IBM® SPSS® Modeler. L'accesso si effettua dal gruppo di programmi IBM® SPSS® Modeler nel menu Start di Windows. Il file *telco_custcat.str* si trova nella directory *streams*.

Questo esempio si concentra sull'utilizzo dei dati demografici per prevedere gli schemi di utilizzo dei servizi. Il campo obiettivo *catcli* ha quattro valori possibili, che corrispondono ai quattro gruppi di clienti, come segue:

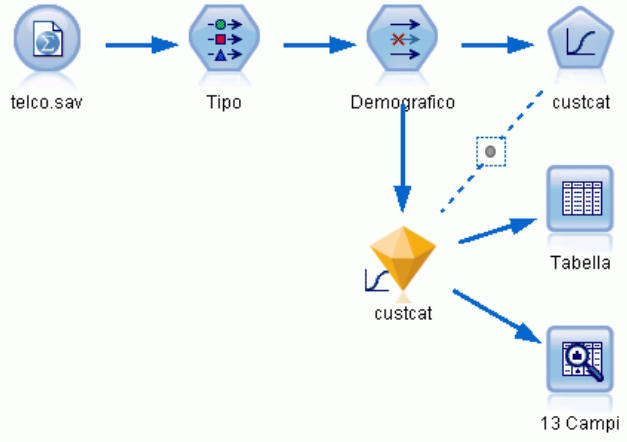
Valore	Label
1	Servizio Base
2	E-Service
3	Servizio Plus
4	Servizio Totale

Poiché l'obiettivo ha più categorie, viene utilizzato un modello multinomiale. Nel caso di un obiettivo che presenta due categorie distinte, per esempio sì/no, vero/falso o abbandono/non abbandono, è possibile creare invece un modello binomiale. Per ulteriori informazioni, vedere l'argomento Tasso di abbandono nelle telecomunicazioni (Regressione logistica binomiale) in il capitolo 13 a pag. 160.

Creazione dello stream

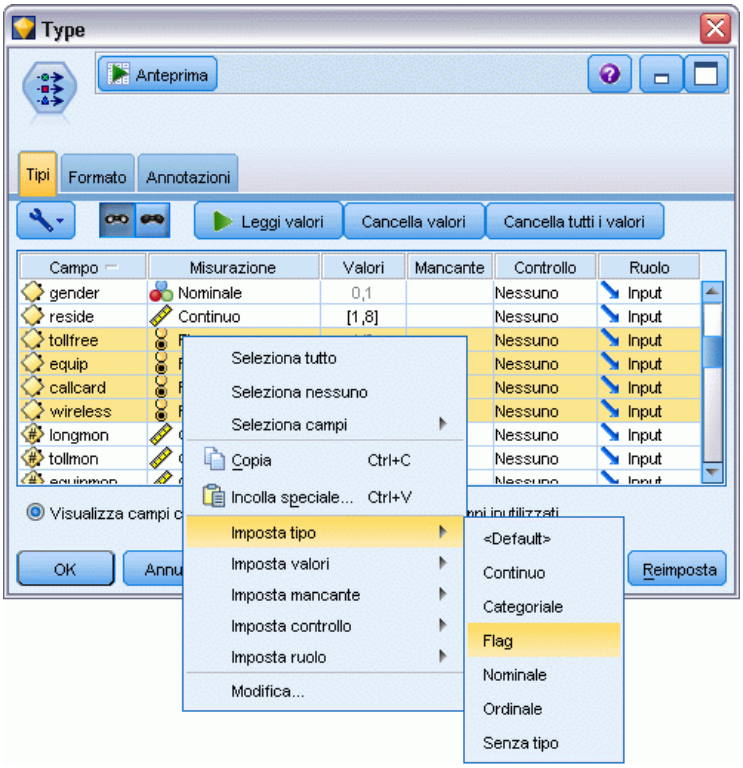
- Aggiungere un nodo di input File Statistics che punta a *telco.sav* nella cartella *Demos*.

Figura 12-1
Stream campione per la classificazione dei clienti mediante regressione logistica multinomiale



- Aggiungere un nodo Tipo e fare clic su Leggi valori, assicurandosi che tutti i livelli di misurazione siano impostati correttamente. Per esempio, quasi tutti i campi con valore 0 e 1 si possono considerare flag.

Figura 12-2
Impostazione del livello di misurazione per più campi

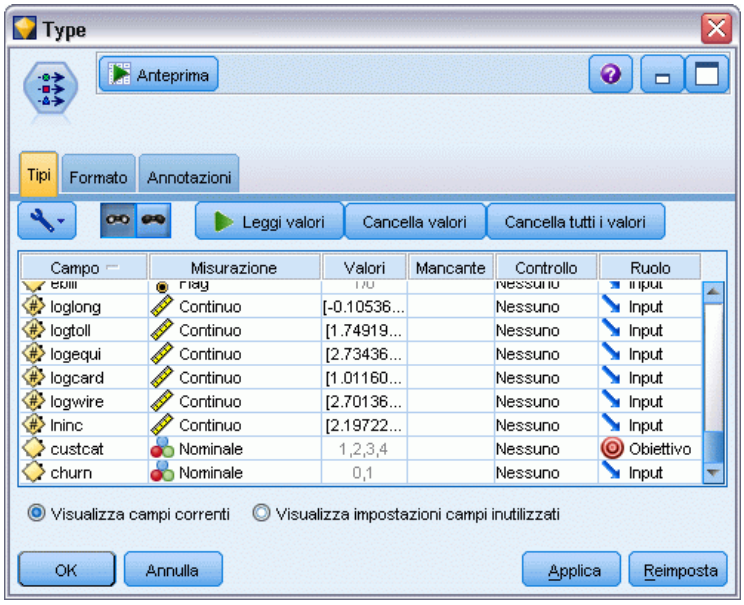


Suggerimento: se si desidera modificare le proprietà di più campi con valori simili (quali 0/1), fare clic sull'intestazione della colonna *Valori* per ordinare i campi in base al valore e tenere premuto il tasto Maiusc mentre si selezionano con il mouse o i tasti freccia tutti i campi da modificare. A questo punto, è possibile fare clic con il pulsante destro del mouse sulla selezione per modificare il livello di misurazione o altri attributi dei campi selezionati.

Si noti che *sesso* è considerato più correttamente un campo con un insieme di due valori, anziché un flag; pertanto, lasciare il suo valore Misurazione impostato su Nominale.

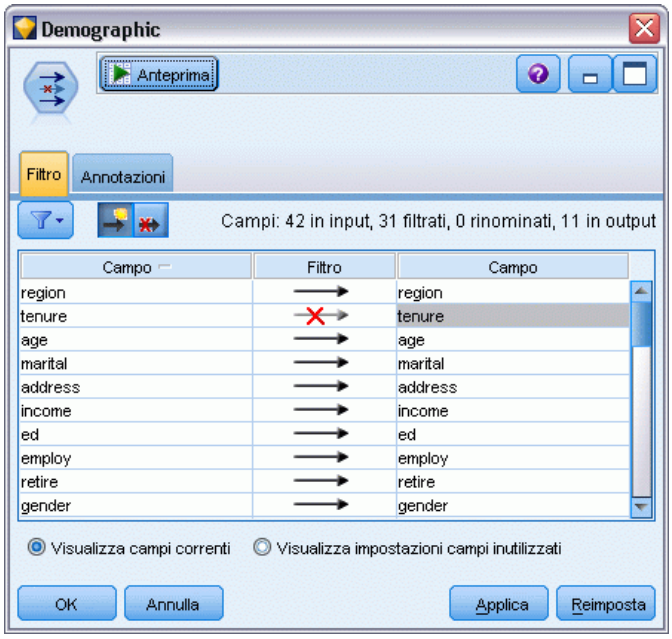
- Impostare il ruolo del campo *catcli* su Obiettivo. Tutti gli altri campi dovrebbero avere il ruolo impostato su Input.

Figura 12-3
Impostazione del ruolo del campo



Poiché questo esempio prende in esame dati demografici, utilizzare il nodo Filtro per includere solo i campi rilevanti (*regione*, *età*, *statciv*, *indirizzo*, *reddito*, *istruz*, *impiego*, *pensionato/a*, *sesso*, *residenza* e *catpers*). Ai fini di questa analisi, è possibile escludere gli altri campi.

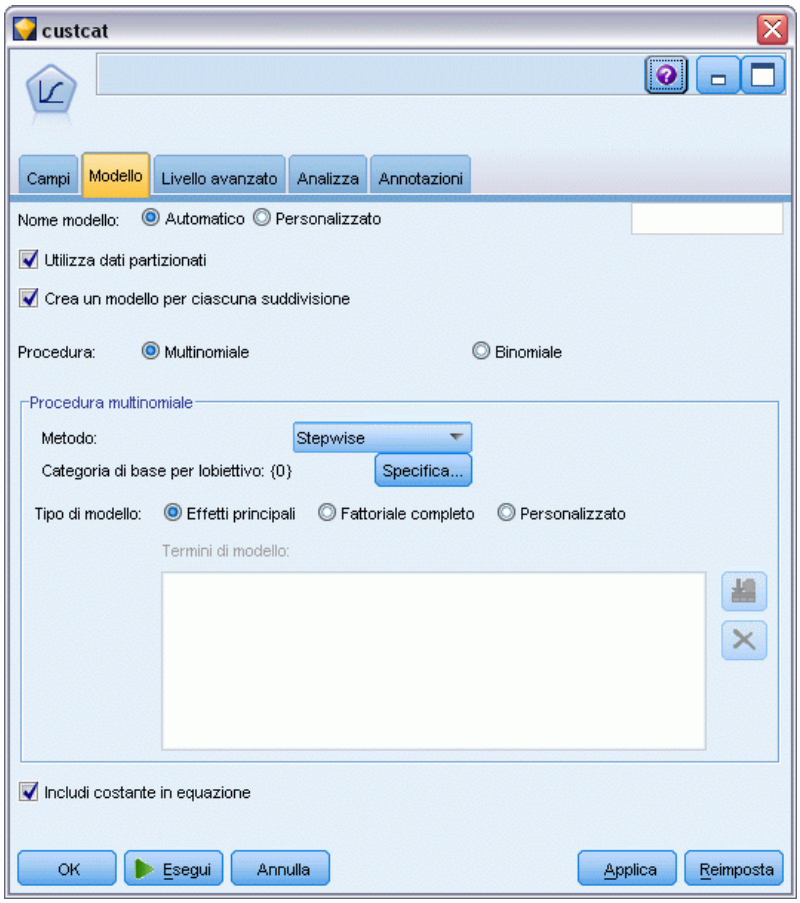
Figura 12-4
Filtro di campi demografici



In alternativa, anziché escludere tali campi, è possibile modificarne il ruolo in Nessuno oppure selezionare i campi da utilizzare nel nodo Modelli.

- Nel nodo Logistica, fare clic sulla scheda Modello e selezionare il metodo Stepwise. Selezionare anche Multinomiale, Effetti principali e Includi costante in equazione.

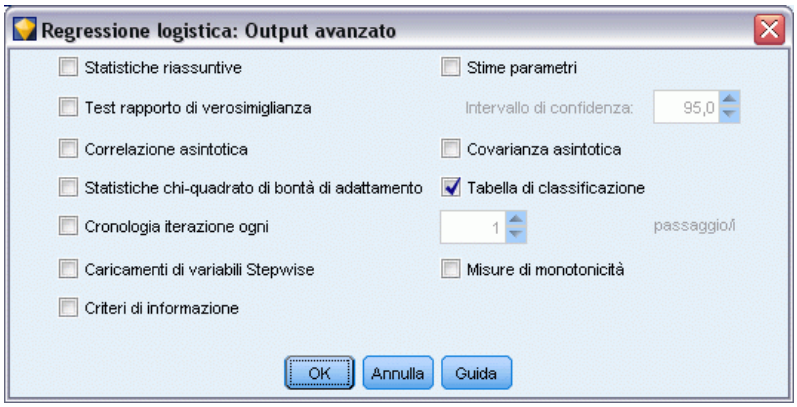
Figura 12-5
Scelta delle opzioni del modello



Mantenere il valore 1 per la categoria Base. Il modello eseguirà il confronto fra gli altri clienti e i clienti che hanno sottoscritto il Servizio Base.

- Nella scheda Livello avanzato, selezionare la modalità Livello avanzato e Output, quindi, nella finestra di dialogo Output avanzato, selezionare Tabella di classificazione.

Figura 12-6
Scelta delle opzioni di output

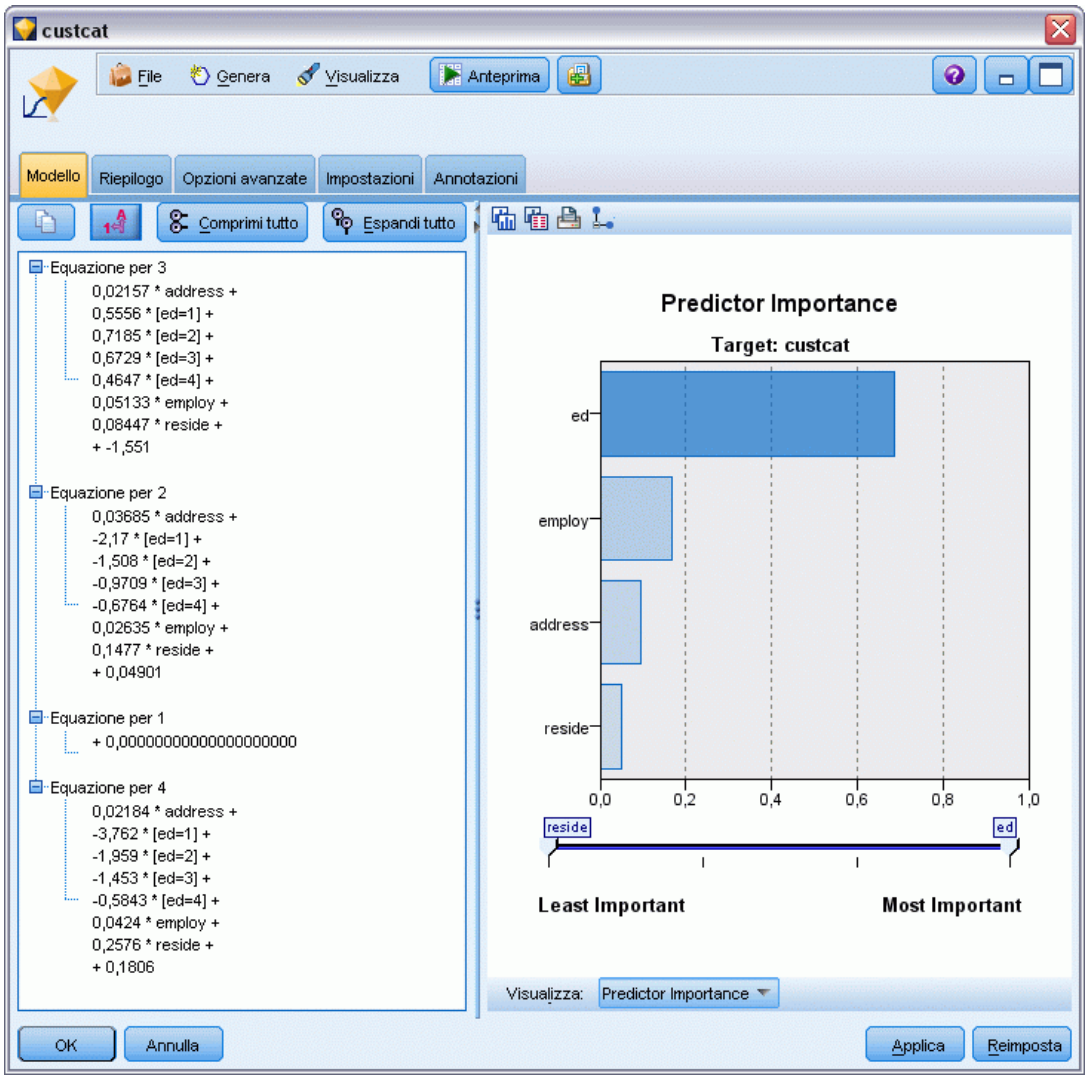


Visualizzazione del modello

- Eseguire il nodo per generare il modello, che viene aggiunto al file palette Modelli nell'angolo superiore destro. Per visualizzare i relativi dettagli, fare clic con il pulsante destro del mouse sul nodo del modello generato e selezionare Visualizza.

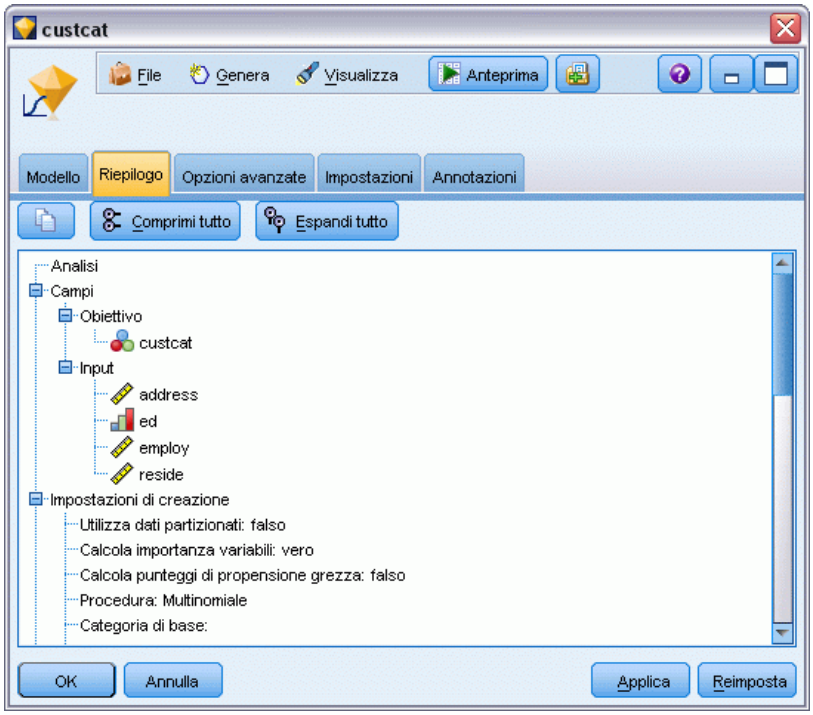
La scheda Modello visualizza le equazioni utilizzate per assegnare i record alle singole categorie del campo obiettivo. Esistono quattro possibili categorie, una delle quali è la categoria Base per la quale non vengono mostrati i dettagli dell'equazione. I dettagli vengono invece mostrati per le tre equazioni rimanenti, dove la categoria 3 rappresenta il Servizio Plus, e così via.

Figura 12-7
Visualizzazione dei risultati del modello



La scheda Riepilogo, fra le altre cose, mostra i campi obiettivo e di input (predittori) utilizzati dal modello. Si noti che questi sono i campi che sono stati scelti in base al metodo Stepwise e non rappresentano l'elenco completo sottoposto all'analisi.

Figura 12-8
Riepilogo del modello che mostra i campi obiettivo e di input

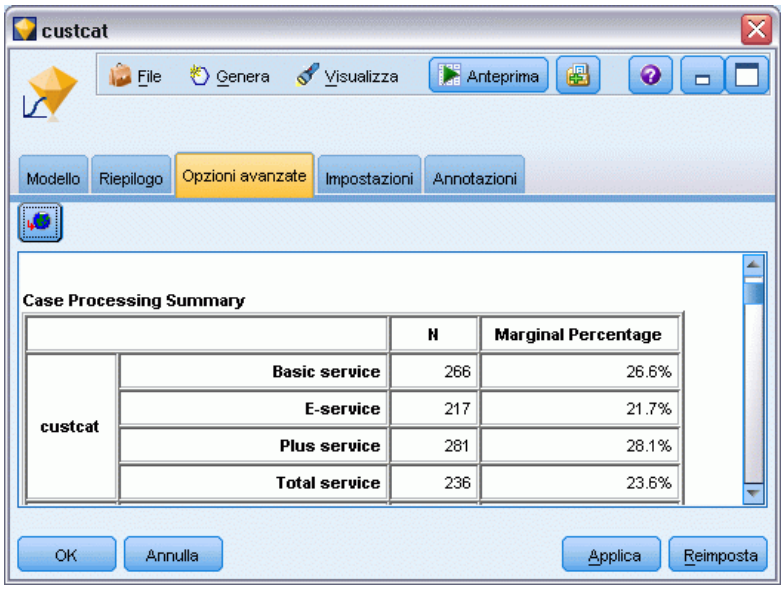


Gli elementi visualizzati nella scheda Opzioni avanzate dipendono dalle opzioni selezionate nella finestra di dialogo Output avanzato nel nodo Modelli.

Un elemento che è sempre presente è il Riepilogo dell'elaborazione del caso, che mostra la percentuale dei record che rientra nelle singole categorie del campo obiettivo. In questo modo, si dispone di un modello nullo da utilizzare come base per il confronto.

Se non si crea un modello che utilizza i predittori, il sistema migliore consiste nell'assegnare tutti i clienti al gruppo più comune, vale a dire quello del Servizio Plus.

Figura 12-9
Riepilogo dell'elaborazione dei casi

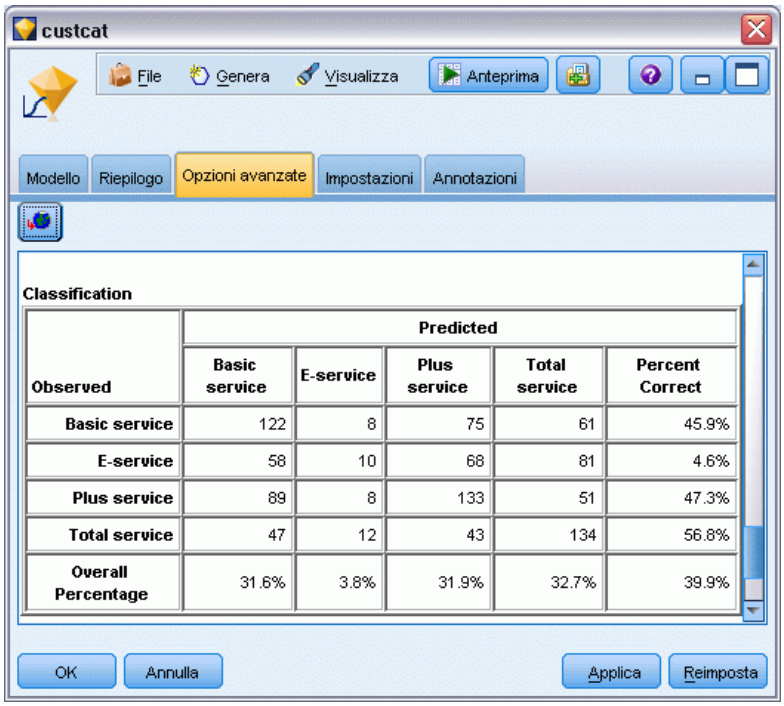


In base ai dati di addestramento, se tutti i clienti venissero assegnati al modello nullo, si otterrebbe un risultato corretto 281 volte su 1000, vale a dire, il 28.1% delle volte. La scheda Opzioni avanzate contiene ulteriori informazioni che consentono di analizzare le previsioni del modello. A questo punto è possibile confrontare le previsioni con i risultati del modello nullo per verificare l'efficacia del modello con i propri dati.

In fondo alla scheda Opzioni avanzate, la Tabella di classificazione mostra i risultati relativi al proprio modello, che risulta corretto il 39.9% delle volte.

In particolare, il modello riporta ottimi risultati nell'identificazione dei clienti del Servizio Totale (categoria 4), ma è decisamente inefficace nell'identificare i clienti E-Service (categoria 2). Per ottenere un grado di precisione più elevato relativamente ai clienti della categoria 2, potrebbe essere necessario trovare un altro predittore per la loro identificazione.

Figura 12-10
Tabella classificazioni



Classification

Observed	Predicted				
	Basic service	E-service	Plus service	Total service	Percent Correct
Basic service	122	8	75	61	45.9%
E-service	58	10	68	81	4.6%
Plus service	89	8	133	51	47.3%
Total service	47	12	43	134	56.8%
Overall Percentage	31.6%	3.8%	31.9%	32.7%	39.9%

A seconda di ciò che si vuole prevedere, il modello potrebbe essere perfettamente idoneo alle proprie esigenze. Per esempio, se non si è interessati all'identificazione dei clienti nella categoria 2, questo modello può considerarsi sufficientemente preciso, come nel caso in cui E-service fosse un cosiddetto "loss-leader", che genera bassi profitti.

Se, per esempio, il rendimento sul capitale investito più elevato deriva dai clienti delle categorie 3 o 4, questo modello può fornire le informazioni necessarie.

Per valutare la misura dell'adattamento effettivo del modello ai dati, durante la creazione del modello sono disponibili una serie di funzioni di diagnostica nella finestra di dialogo Output avanzato. Per ulteriori informazioni, vedere l'argomento Output avanzato dell'insieme di modelli Logistica in il capitolo 10 in *IBM SPSS Modeler 15 Nodi Modelli*. Per una spiegazione dei fondamenti matematici dei metodi di modellazione utilizzati in IBM® SPSS® Modeler, vedere il manuale *SPSS Modeler Algorithms Guide*, disponibile nella directory *\Documentation* del disco di installazione.

Si noti inoltre che questi risultati si basano soltanto sui dati di addestramento. Per un'indicazione dell'efficacia con cui il modello potrà essere esteso ad altri dati in una situazione reale, utilizzare un nodo Partizione per estrarre un sottoinsieme di record da utilizzare per test e validazione. Per ulteriori informazioni, vedere l'argomento Nodo Partizione in il capitolo 4 in *IBM SPSS Modeler 15 Nodi di input, elaborazione e output*.

Capitolo
13

Tasso di abbandono nelle telecomunicazioni (Regressione logistica binomiale)

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

In questo esempio viene utilizzato lo stream denominato *telco_churn.str*, che fa riferimento al file di dati denominato *telco.sav*. Questi file sono disponibili nella directory *Demos* di tutte le installazioni di IBM® SPSS® Modeler. L'accesso si effettua dal gruppo di programmi IBM® SPSS® Modeler nel menu Start di Windows. Il file *telco_churn.str* si trova nella directory *streams*.

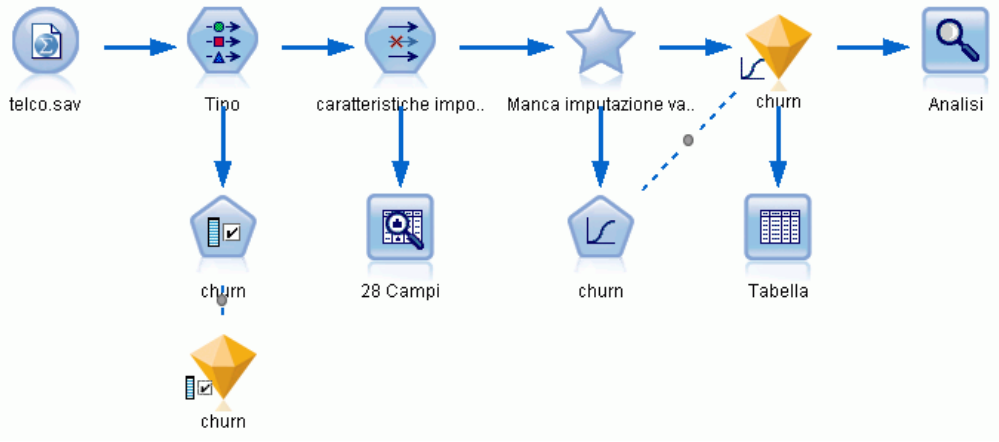
Si supponga, per esempio, che una società che fornisce servizi di telecomunicazioni sia preoccupata del numero dei clienti persi a favore della concorrenza. Se, mediante i dati di utilizzo del servizio, è possibile prevedere quali clienti potrebbero passare a un altro provider, sarà possibile personalizzare le offerte in modo da riuscire a trattenere il maggior numero di clienti possibile.

Questo esempio si concentra sull'impiego dei dati di utilizzo per prevedere la perdita di clienti (tasso di abbandono). Poiché l'obiettivo ha due categorie distinte, viene adottato un modello binomiale. Se l'obiettivo avesse più categorie, sarebbe possibile creare un modello multinomiale.
Per ulteriori informazioni, vedere l'argomento Classificazione dei clienti nelle telecomunicazioni (Regressione logistica multinomiale) in il capitolo 12 a pag. 150.

Creazione dello stream

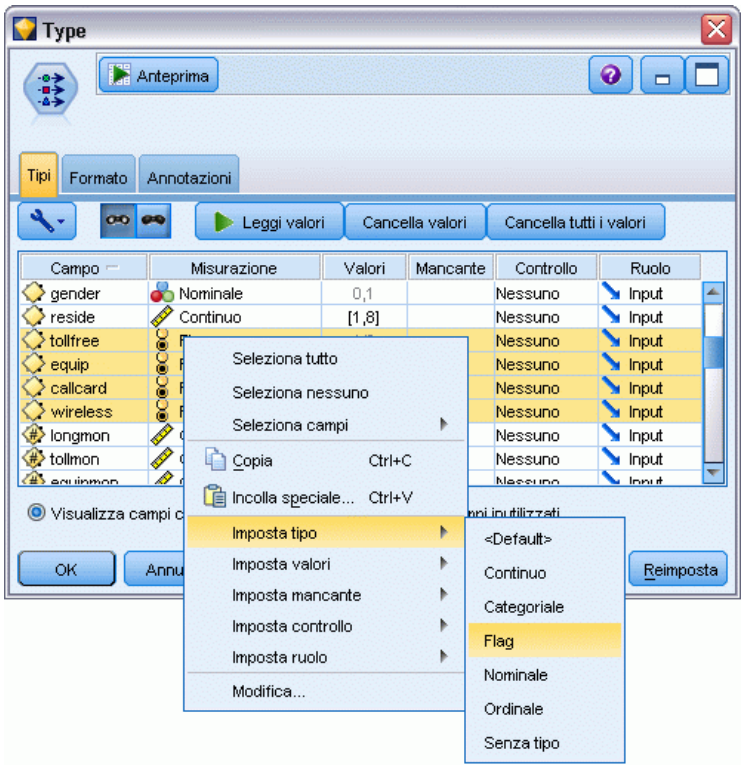
- Aggiungere un nodo di input File Statistics che punta a *telco.sav* nella cartella *Demos*.

Figura 13-1
Stream campione per la classificazione dei clienti mediante regressione logistica binomiale



- Aggiungere un nodo Tipo per definire i campi, assicurandosi che tutti i livelli di misurazione siano impostati correttamente. Per esempio, la maggior parte dei campi con valore 0 e 1 può essere considerata di tipo flag, ma alcuni campi, come quello relativo al sesso, vengono visualizzati più accuratamente come campi nominali con due valori.

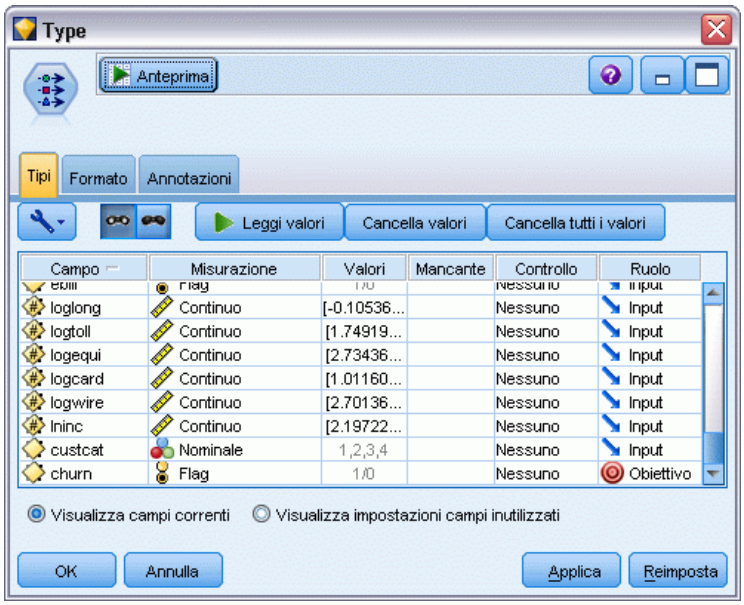
Figura 13-2
Impostazione del livello di misurazione per più campi



Suggerimento: se si desidera modificare le proprietà di più campi con valori simili (quali 0/1), fare clic sull'intestazione della colonna *Valori* per ordinare i campi in base al valore e tenere premuto il tasto Maiusc mentre si selezionano con il mouse o i tasti freccia tutti i campi da modificare. A questo punto, è possibile fare clic con il pulsante destro del mouse sulla selezione per modificare il livello di misurazione o altri attributi dei campi selezionati.

- Impostare il livello di misurazione per il campo *Tasso di abbandono* su Flag e il ruolo su Obiettivo. Tutti gli altri campi dovrebbero avere il ruolo impostato su Input.

Figura 13-3
Impostazione del livello di misurazione e del ruolo per il campo Tasso di abbandono



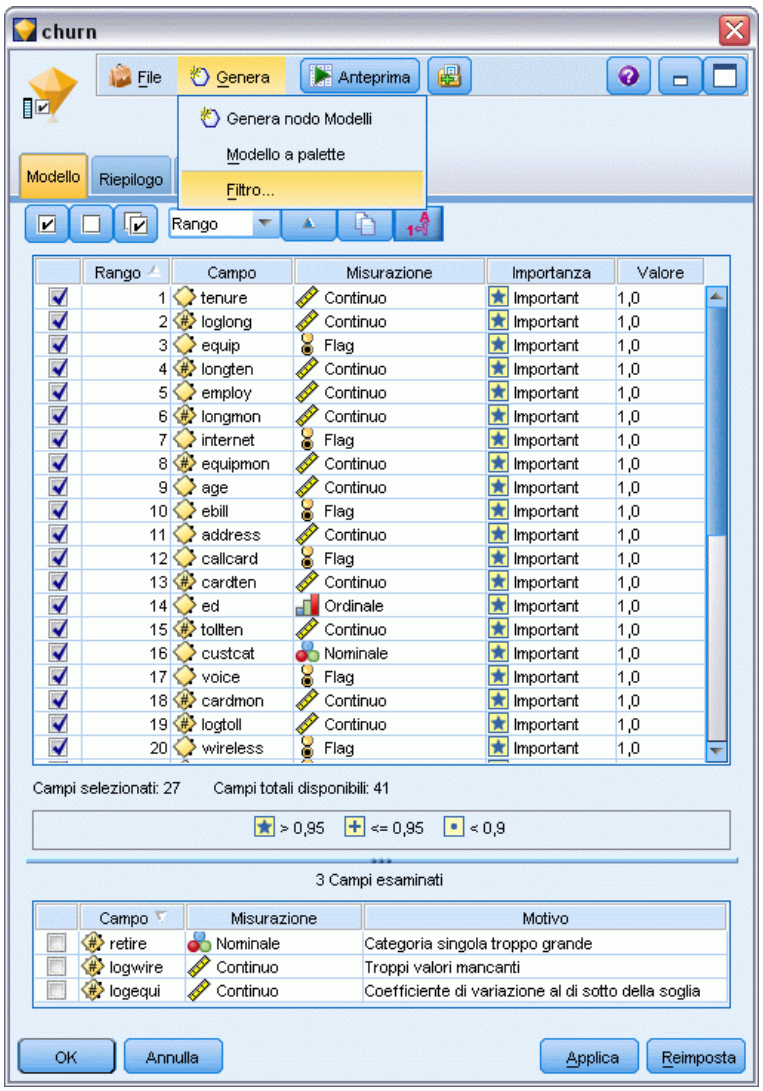
- Aggiungere un nodo Modelli Selezione funzioni al nodo Tipo.

  L'utilizzo del nodo Selezione funzioni consente di rimuovere i predittori o i dati che non aggiungono informazioni utili rispetto alla relazione predittore/obiettivo.

- Esecuzione dello stream.

▶ Aprire l'insieme di modelli risultante e, dal menu Genera, scegliere Filtro per creare un nodo Filtro.

Figura 13-4
Generazione di un nodo Filtro da un nodo Selezione funzioni



Rango	Campo	Misurazione	Importanza	Valore
1	tenure	Continuo	Important	1,0
2	loglong	Continuo	Important	1,0
3	equip	Flag	Important	1,0
4	longten	Continuo	Important	1,0
5	employ	Continuo	Important	1,0
6	longmon	Continuo	Important	1,0
7	internet	Flag	Important	1,0
8	equipmon	Continuo	Important	1,0
9	age	Continuo	Important	1,0
10	ebill	Flag	Important	1,0
11	address	Continuo	Important	1,0
12	callcard	Flag	Important	1,0
13	cardten	Continuo	Important	1,0
14	ed	Ordinale	Important	1,0
15	tollten	Continuo	Important	1,0
16	custcat	Nominale	Important	1,0
17	voice	Flag	Important	1,0
18	cardmon	Continuo	Important	1,0
19	logtoll	Continuo	Important	1,0
20	wireless	Flag	Important	1,0

Campo	Misurazione	Motivo
retire	Nominale	Categoria singola troppo grande
logwire	Continuo	Troppi valori mancanti
logequi	Continuo	Coefficiente di variazione al di sotto della soglia

Non tutti i dati contenuti nel file *telco.sav* risulteranno utili per la previsione del tasso di abbandono. È possibile utilizzare il filtro per selezionare solo i dati considerati importanti come predittori.

▶ Nella finestra di dialogo Genera filtro, selezionare Tutti i campi contrassegnati: Importante e fare clic su OK.

▶ Collegare il nodo Filtro generato al nodo Tipo.

Figura 13-5
Selezione di campi importanti

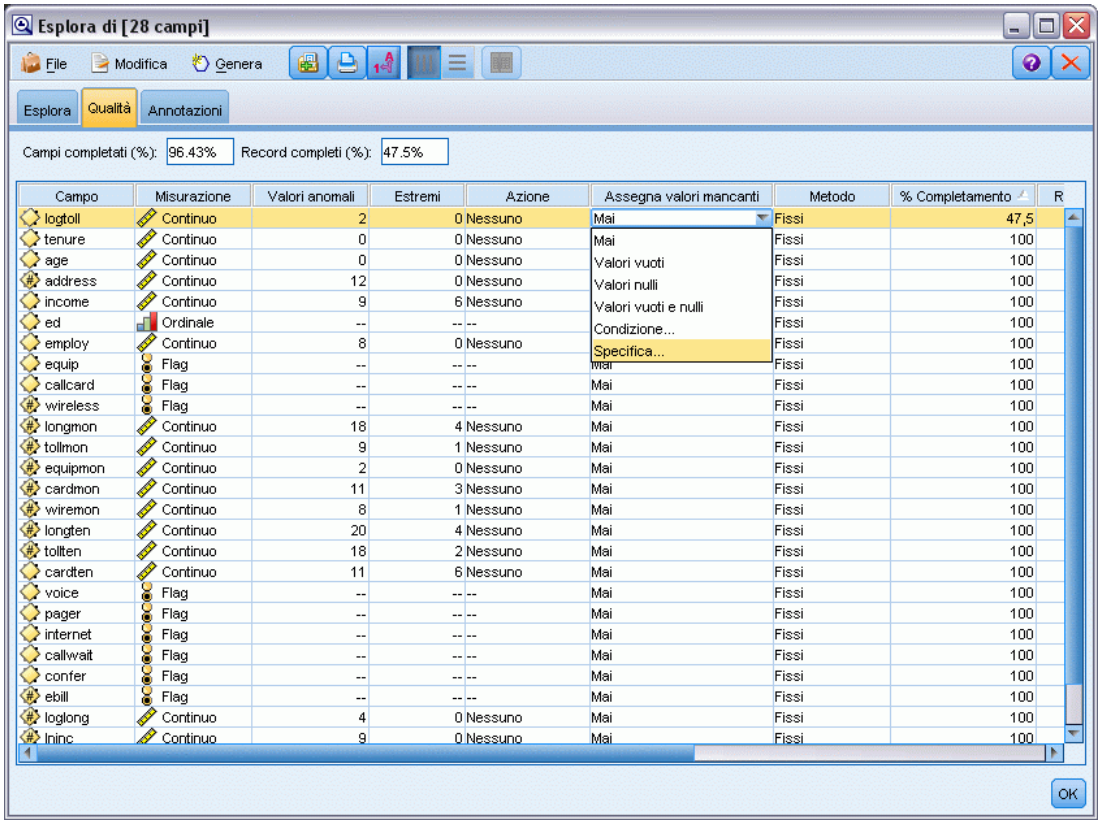
▶ Collegare un nodo Esplora al nodo Filtro generato.

Aprire il nodo Esplora e fare clic su Esegui.

▶ Nella scheda Qualità del browser del nodo Esplora, fare clic sulla colonna *% Complete* per visualizzare la colonna in ordine numerico crescente. Questa operazione consente di identificare i campi con grosse quantità di dati mancanti; in questo caso, l'unico campo che è necessario modificare è *registra_costo*, che ha una percentuale di completezza inferiore al 50%.

▶ Nella colonna *Assegna valori mancanti* per *registra_costo*, fare clic su Specifica.

Figura 13-6
Assegnazione dei valori mancanti per registra costo



▶ Per Assegna quando, selezionare Valori vuoti e nulli. Per Fisso come, selezionare Media e fare clic su OK.

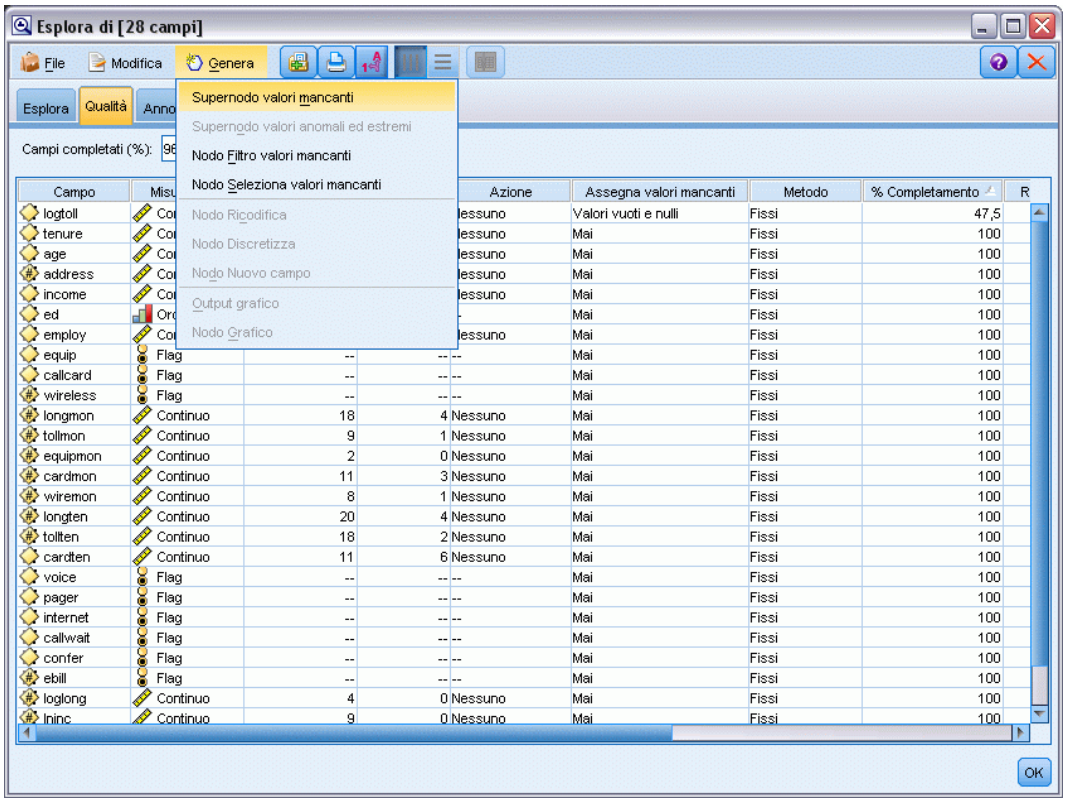
La selezione di Media fa sì che i valori assegnati non abbiano un impatto negativo sulla media di tutti i valori dei dati globali.

Figura 13-7
Selezione delle impostazioni di assegnazione

- Nella scheda Qualità del browser del nodo Esplora, generare il Supernodo valori mancanti. Per farlo, dai menu scegliere:

  Genera > Supernodo valori mancanti

Figura 13-8
Generazione di un Supernodo valori mancanti



Nella finestra di dialogo Supernodo valori mancanti, incrementare a 50% il valore di Dimensione campione e fare clic su OK.

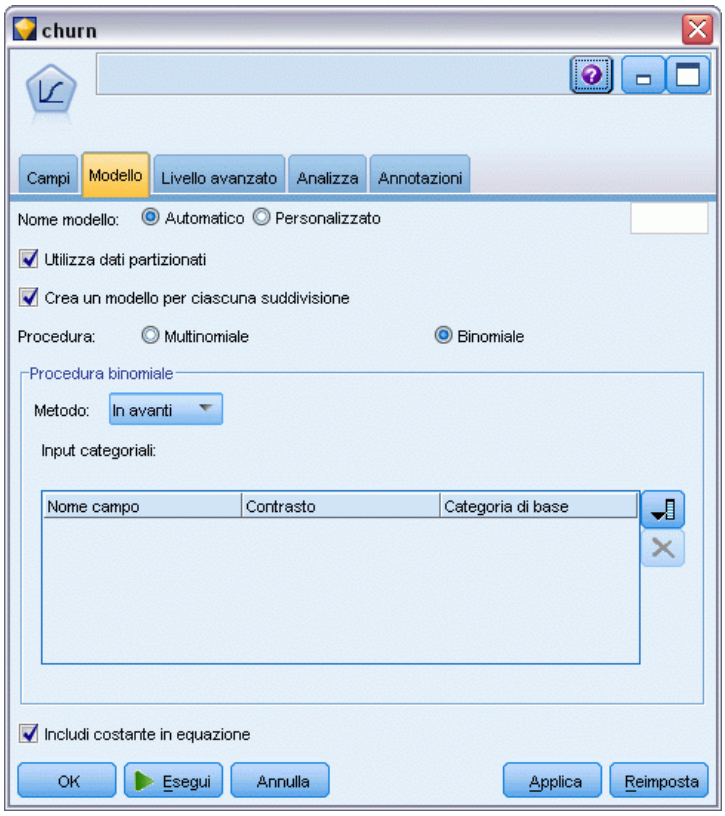
Il Supernodo viene visualizzato nell'area di disegno dello stream, con il titolo: *Assegnazione valori mancanti*.

- Collegare il Supernodo al nodo Filtro.

Figura 13-9
Specifica della dimensione del campione

- Aggiungere un nodo Logistica al Supernodo.
- Nel nodo Logistica, fare clic sulla scheda Modello e selezionare la procedura Binomiale. Nell'area della *procedura binomiale*, selezionare il metodo In avanti.

Figura 13-10
Scelta delle opzioni del modello



- Nella scheda Livello avanzato, selezionare la modalità Livello avanzato e fare clic su Output. Viene visualizzata la finestra di dialogo Output avanzato.
- Nella finestra di dialogo Output avanzato, selezionare A ogni passaggio come tipo del campo *Visualizza*. Selezionare Cronologia iterazione e Stime parametri e fare clic su OK.

Figura 13-11
Scelta delle opzioni di output

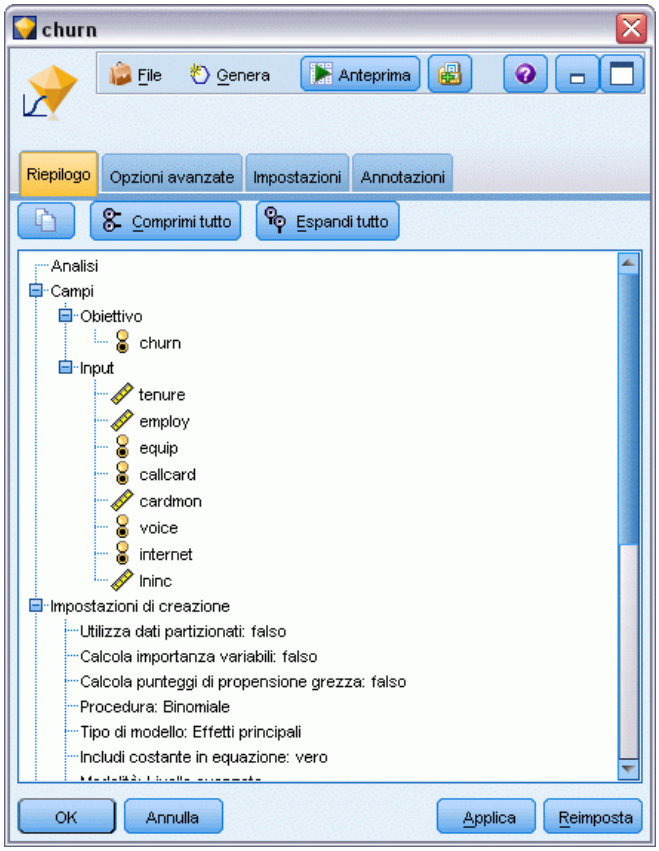
Visualizzazione del modello

- Nel nodo Logistica, fare clic su Esegui per creare il modello.

L'insieme di modelli viene aggiunto all'area di disegno dello stream e alla palette Modelli nell'angolo superiore destro. Per visualizzarne i dettagli, fare clic con il pulsante destro del mouse sull'insieme di modelli e selezionare Modifica o Sfoglia.

La scheda Riepilogo, fra le altre cose, mostra i campi obiettivo e di input (predittori) utilizzati dal modello. Si noti che questi sono i campi che sono stati scelti in base al metodo In avanti e non rappresentano l'elenco completo sottoposto all'analisi.

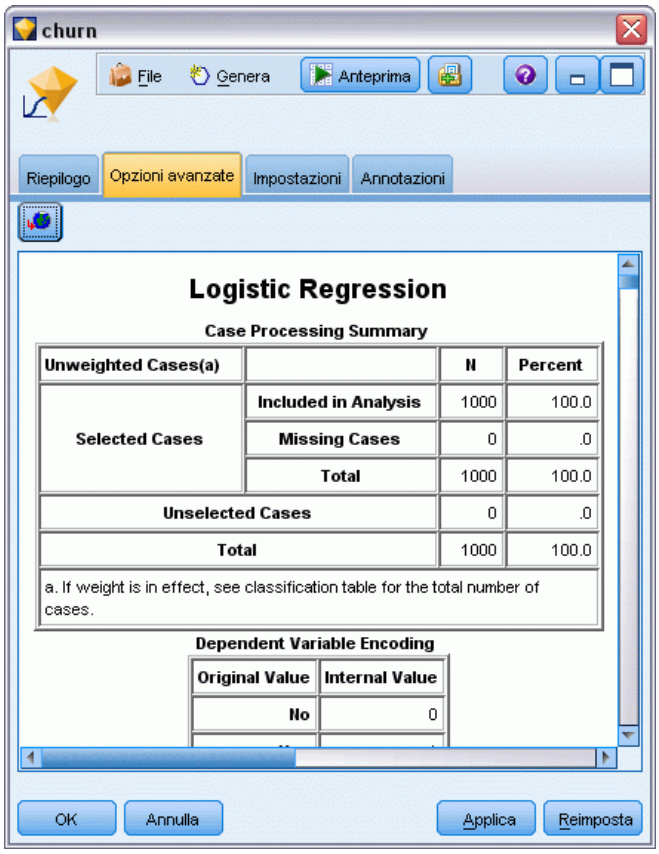
Figura 13-12
Riepilogo del modello che mostra i campi obiettivo e di input



Gli elementi visualizzati nella scheda Opzioni avanzate dipendono dalle opzioni selezionate nella finestra di dialogo Output avanzato nel nodo Logistica. Un elemento che è sempre presente è il Riepilogo dell'elaborazione dei casi, che mostra il numero e la percentuale dei record inclusi

nell'analisi. Inoltre, vengono riportati il numero degli eventuali casi mancanti dove uno o più campi di input non sono disponibili e gli eventuali casi che non sono stati selezionati.

Figura 13-13
Riepilogo dell'elaborazione dei casi

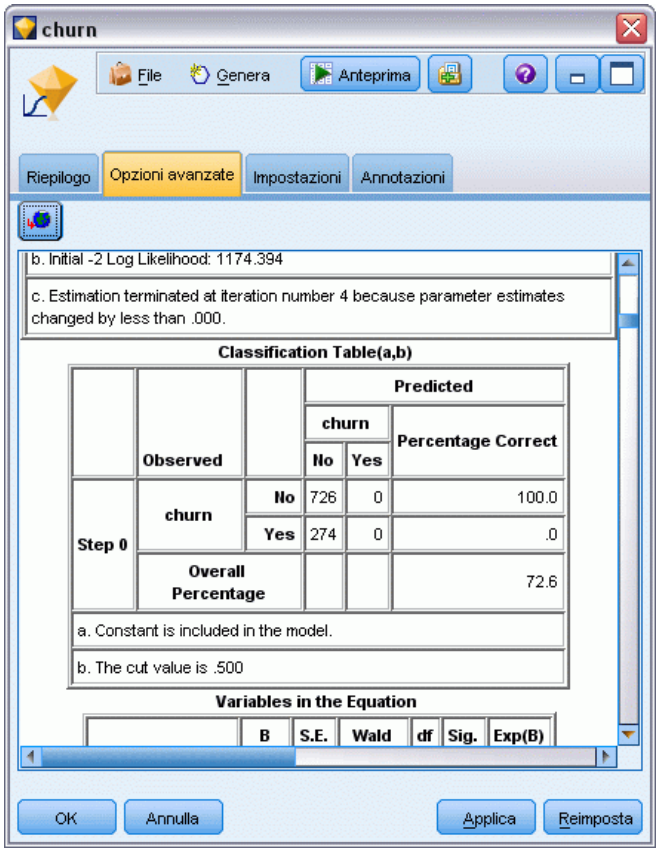


- Scorrere verso il basso fino al Riepilogo dell'elaborazione dei casi per visualizzare la tabella di classificazione sotto Blocco 0: blocco iniziale.

  Il metodo Stepwise in avanti inizia con un modello nullo, vale a dire un modello senza predittori, che può essere utilizzato come base per il confronto con il modello creato finale. Il modello nullo, per convenzione, prevede tutto sotto forma di 0; pertanto, tale modello è accurato al 72.6%, semplicemente perché i 726 clienti che non sono passati a un altro gestore sono stati previsti

correttamente. Tuttavia, i clienti che sono passati a un altro gestore non sono stati previsti correttamente.

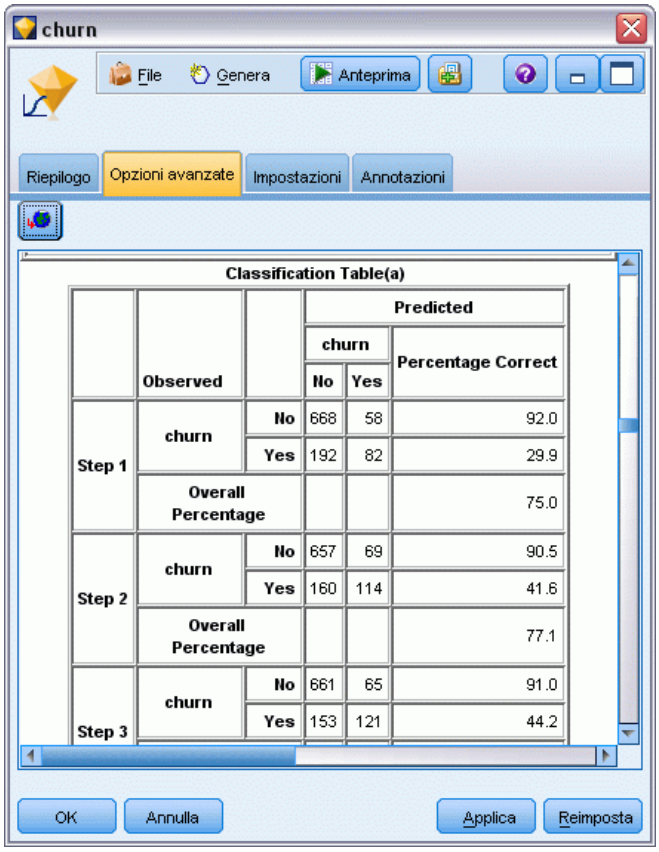
Figura 13-14
Tabella di classificazione iniziale - Blocco 0



- Scorrere verso il basso per visualizzare la tabella di classificazione sotto Blocco 1: Metodo = Stepwise in avanti.

Questa tabella di classificazione mostra i risultati del modello ogni volta che un predittore viene aggiunto in corrispondenza dei singoli passi. Già nel primo passo, dopo che è stato utilizzato un solo predittore, il modello incrementa la precisione della previsione del tasso di abbandono da 0.0% a 29.9%

Figura 13-15
Tabella di classificazione - Blocco 1



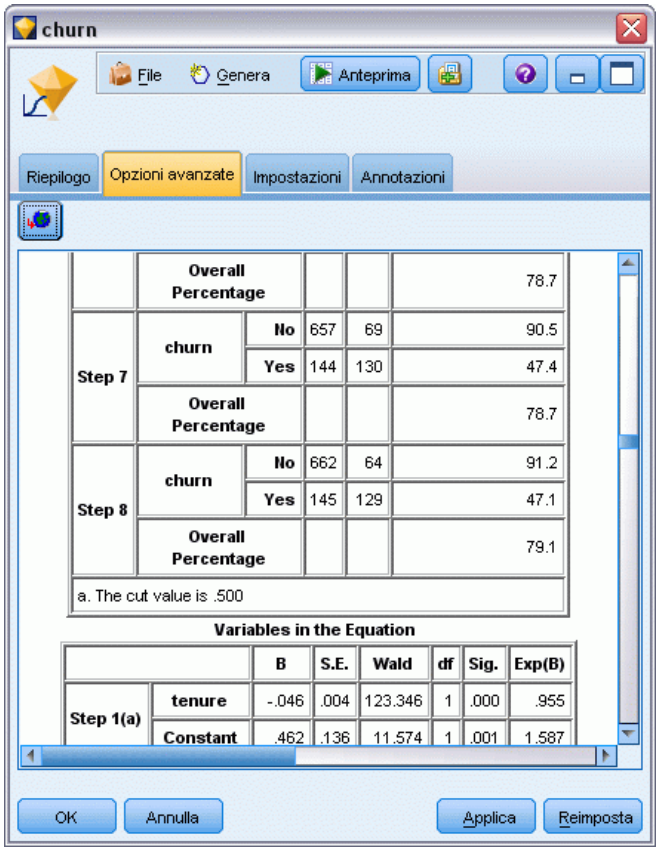
	Observed		Predicted		
			churn		Percentage Correct
			No	Yes	
Step 1	churn	No	668	58	92.0
		Yes	192	82	29.9
	Overall Percentage				75.0
Step 2	churn	No	657	69	90.5
		Yes	160	114	41.6
	Overall Percentage				77.1
Step 3	churn	No	661	65	91.0
		Yes	153	121	44.2

▶ Scorrere verso il basso fino alla fine della Tabella di classificazione.

La tabella di classificazione mostra che il passo 8 è l'ultimo passo. A questo punto, l'algoritmo ha deciso che non è necessario aggiungere altri predittori al modello. Anche se il grado di precisione dei clienti che non passano a un altro gestore è lievemente diminuito, passando al 91,2%, il grado di precisione della previsione dei clienti che invece hanno abbandonato il gestore è passato

dallo 0% iniziale al 47.1%. Si tratta di un miglioramento significativo rispetto al modello nullo originale che non utilizzava predittori.

Figura 13-16
Tabella di classificazione - Blocco 1



Per un cliente che desidera ridurre il tasso di abbandono, riuscire quasi a dimezzare tale tasso rappresenta già un enorme passo in avanti nella protezione delle proprie entrate.

Nota: questo esempio mostra anche che considerare la percentuale complessiva come indicazione della precisione di un modello può, in alcuni casi, essere fuorviante. Il modello nullo originale aveva un tasso di precisione complessivo pari a 72.6%, mentre il modello finale previsto è accurato al 79,1%; tuttavia, come si è visto, il grado di precisione delle previsioni effettive relative alle categorie individuali era decisamente diverso.

Per valutare la misura dell'adattamento effettivo del modello ai dati, durante la creazione del modello sono disponibili una serie di funzioni di diagnostica nella finestra di dialogo Output avanzato. Per ulteriori informazioni, vedere l'argomento Output avanzato dell'insieme di modelli Logistica in il capitolo 10 in *IBM SPSS Modeler 15 Nodi Modelli*. Per una spiegazione dei fondamenti matematici dei metodi di modellazione utilizzati in IBM® SPSS® Modeler, vedere il manuale *SPSS Modeler Algorithms Guide*, disponibile nella directory *\Documentation* del disco di installazione.

Si noti inoltre che questi risultati si basano soltanto sui dati di addestramento. Per un'indicazione dell'efficacia con cui il modello potrà essere esteso ad altri dati in una situazione reale, utilizzare un nodo Partizione per estrarre un sottoinsieme di record da utilizzare per test e validazione. Per ulteriori informazioni, vedere l'argomento Nodo Partizione in il capitolo 4 in *IBM SPSS Modeler 15 Nodi di input, elaborazione e output*.

Capitolo
14

Previsione dell'utilizzo della larghezza di banda (Serie storica)

Previsione mediante il nodo Serie storica

Un analista di un provider nazionale di servizi a banda larga deve creare delle previsioni sugli abbonamenti in modo da poter stimare l'uso dei servizi a banda larga. È necessario effettuare una previsione per ognuno dei mercati locali che costituiscono la base nazionale degli abbonati. Per generare previsioni relative ai dati dei tre mesi successivi per una serie di mercati locali si utilizzerà la modellazione di serie storiche. Un secondo esempio mostra come si possono convertire i dati di input se non hanno un formato adeguato al nodo Serie storica.

In questi esempi viene utilizzato lo stream denominato *broadband_create_models.str*, che fa riferimento al file di dati denominato *broadband_1.sav*. Questi file sono disponibili nella cartella *Demos* di tutte le installazioni di IBM® SPSS® Modeler. L'accesso si effettua dal gruppo di programmi IBM® SPSS® Modeler nel menu Start di Windows. Il file *broadband_create_models.str* si trova nella cartella *streams*.

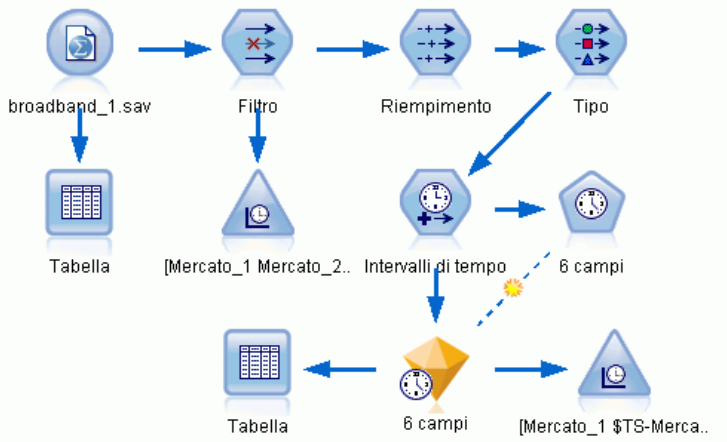
L'ultimo esempio mostra come applicare i modelli salvati a un insieme di dati aggiornato al fine di estendere le previsioni di altri tre mesi.

In SPSS Modeler è possibile produrre più modelli di serie storica in un'unica operazione. Verrà utilizzato un file di origine contenente le serie storiche di 85 mercati diversi, anche se per questione di semplicità verranno creati solo i modelli di cinque mercati e un modello totale per tutti i mercati.

Il file di dati *broadband_1.sav* contiene i dati mensili di utilizzo relativi a tutti gli 85 mercati locali. Ai fini di questo esempio, verranno utilizzate soltanto le prime cinque serie. Verrà creato un modello distinto per ognuna di queste e un modello complessivo per tutte le serie.

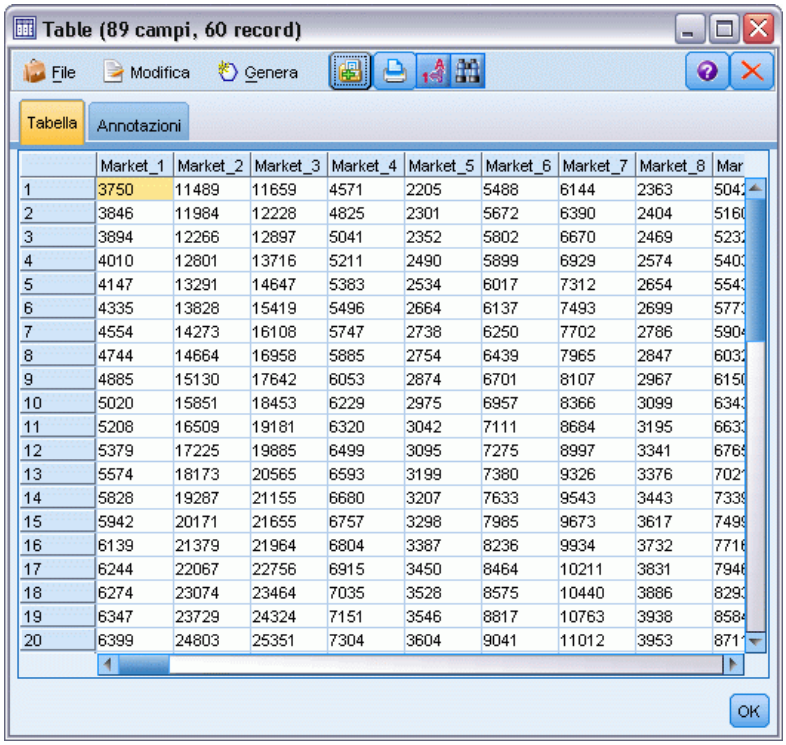
Il file contiene anche un campo data che indica il mese e l'anno per ciascun record. Questo campo verrà utilizzato in un nodo Intervalli di tempo per etichettare i record. Il campo data legge i dati in SPSS Modeler sotto forma di stringa, ma per utilizzarlo in SPSS Modeler sarà necessario convertire il tipo di archiviazione in formato data numerico mediante un nodo Riempimento.

Figura 14-1
Stream campione per mostrare la creazione di modelli di serie storica



Per il nodo Serie storica è necessario che ogni serie si trovi in una colonna separata, con una riga per ogni intervallo. Se necessario, SPSS Modeler fornisce i metodi necessari a trasformare i dati in modo che corrispondano a questo formato.

Figura 14-2
Dati di abbonamento mensili per mercati locali di banda larga



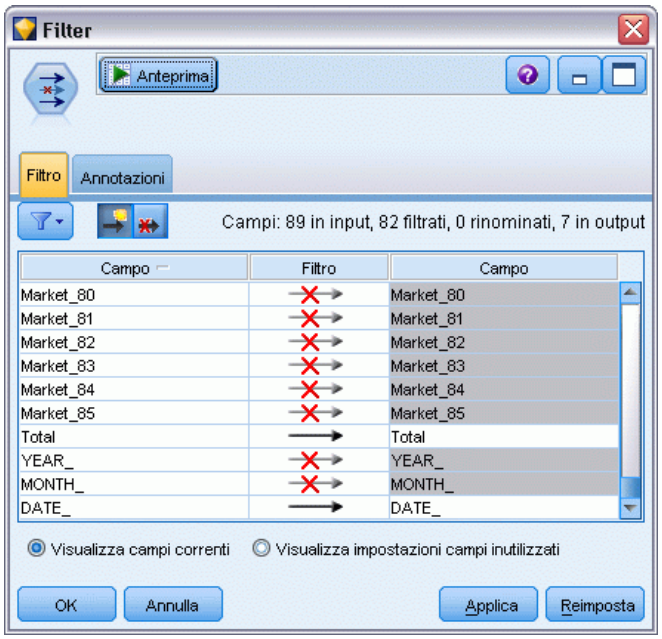
	Market_1	Market_2	Market_3	Market_4	Market_5	Market_6	Market_7	Market_8	Mar
1	3750	11489	11659	4571	2205	5488	6144	2363	504
2	3846	11984	12228	4825	2301	5672	6390	2404	516
3	3894	12266	12897	5041	2352	5802	6670	2469	523
4	4010	12801	13716	5211	2490	5899	6929	2574	540
5	4147	13291	14647	5383	2534	6017	7312	2654	554
6	4335	13828	15419	5496	2664	6137	7493	2699	577
7	4554	14273	16108	5747	2738	6250	7702	2786	590
8	4744	14664	16958	5885	2754	6439	7965	2847	603
9	4885	15130	17642	6053	2874	6701	8107	2967	615
10	5020	15851	18453	6229	2975	6957	8366	3099	634
11	5208	16509	19181	6320	3042	7111	8684	3195	663
12	5379	17225	19885	6499	3095	7275	8997	3341	676
13	5574	18173	20565	6593	3199	7380	9326	3376	702
14	5828	19287	21155	6680	3207	7633	9543	3443	733
15	5942	20171	21655	6757	3298	7985	9673	3617	749
16	6139	21379	21964	6804	3387	8236	9934	3732	771
17	6244	22067	22756	6915	3450	8464	10211	3831	794
18	6274	23074	23464	7035	3528	8575	10440	3886	829
19	6347	23729	24324	7151	3546	8817	10763	3938	858
20	6399	24803	25351	7304	3604	9041	11012	3953	871

Creazione dello stream

- Creare un nuovo stream e aggiungere un nodo di input File Statistics che punti al file *broadband_1.sav*.
- Per semplificare il modello, utilizzare un nodo Filtro per escludere i campi da *Mercato_6* a *Mercato_85* e i campi *MESE_* e *ANNO_*.

Suggerimento: per selezionare più campi adiacenti in un'unica operazione, fare clic sul campo *Mercato_6*, tenere premuto il pulsante sinistro del mouse e trascinare il mouse verso il basso, fino al campo *Mercato_85*. I campi selezionati vengono evidenziati in blu. Per aggiungere altri campi, tenere premuto il tasto Ctrl e fare clic sui campi *MESE_* e *ANNO_*.

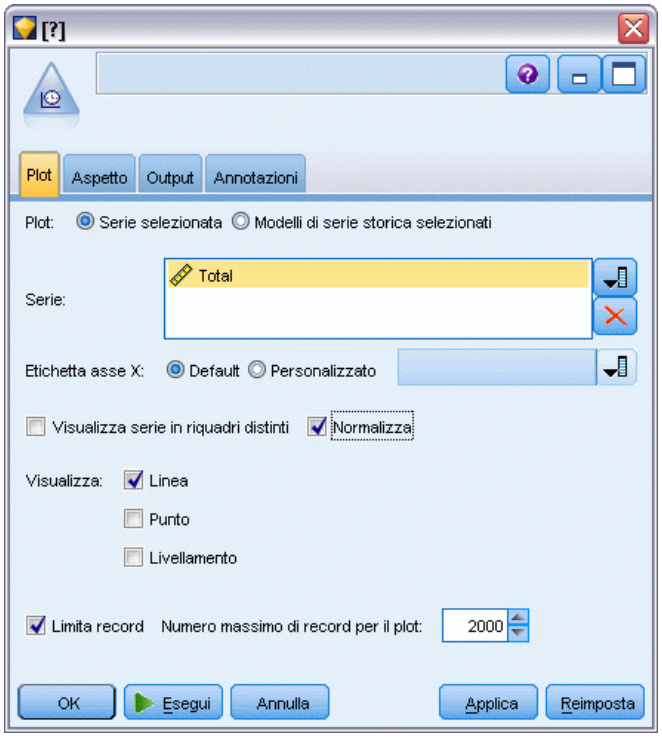
Figura 14-3
Semplificazione del modello



Esame dei dati

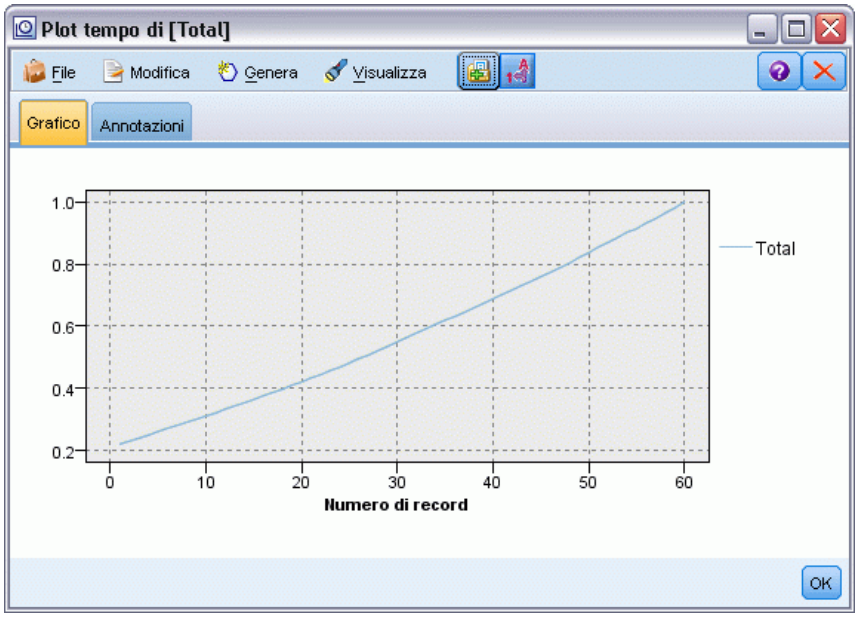
Prima di creare un modello, è sempre utile esaminare dettagliatamente le caratteristiche dei dati usati. I dati evidenziano delle variazioni stagionali? Anche se Expert Modeler può trovare automaticamente il modello stagionale o non stagionale migliore per ogni serie, spesso è possibile ottenere più rapidamente i risultati limitando la ricerca ai modelli non stagionali, se la stagionalità non è presente nei dati. Senza esaminare i dati di tutti i mercati locali, è possibile avere un'idea generale della presenza o dell'assenza di stagionalità rappresentando graficamente il numero totale di abbonati di tutti e cinque i mercati.

Figura 14-4
Rappresentazione grafica del numero totale di abbonati



- Dalla palette Grafici, collegare un nodo Plot tempo al nodo Filtro.
- Aggiungere il campo *Totale* all'elenco Serie.
- Deselezionare le caselle di controllo Visualizza serie in riquadri distinti e Normalizza.
- Fare clic su Esegui.

Figura 14-5
Plot tempo del campo Totale

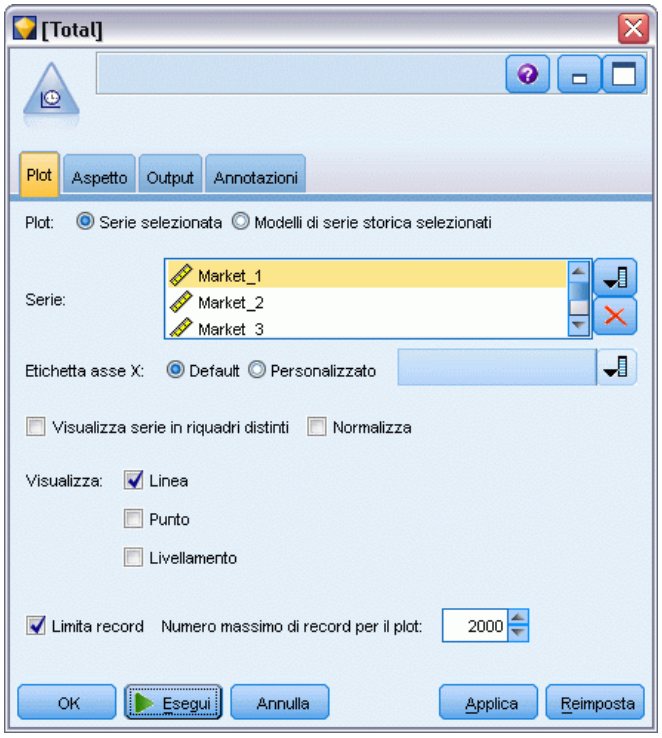


La serie mostra un trend crescente uniforme che non presenta tracce di variazioni stagionali. Anche se non è possibile escludere la presenza di serie con elementi stagionali, la stagionalità non è generalmente una caratteristica significativa dei dati.

Naturalmente, prima di escludere i modelli stagionali è necessario esaminare ciascuna serie, in modo da poterle separare e creare modelli specifici.

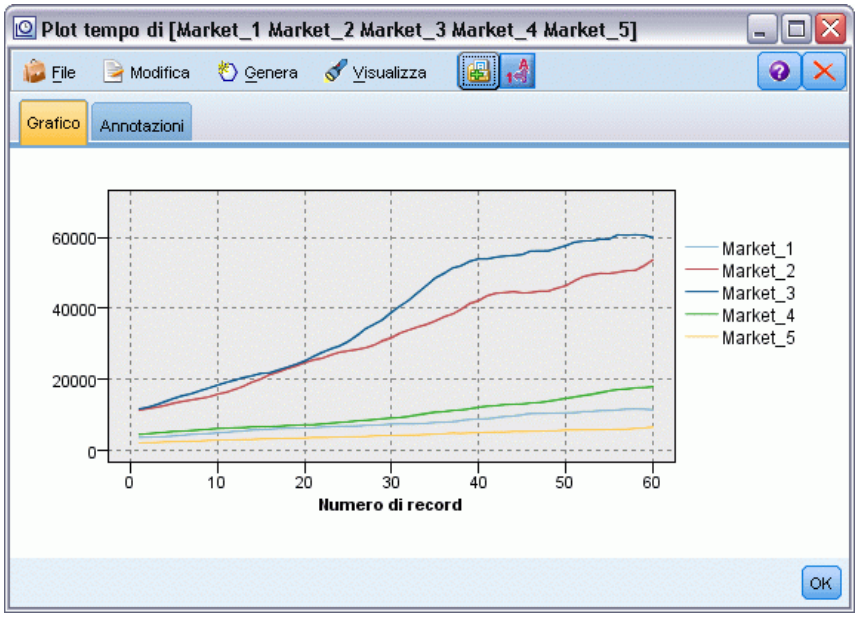
IBM® SPSS® Modeler facilita la rappresentazione grafica di più serie contemporaneamente.

Figura 14-6
Rappresentazione grafica di più serie storiche



- Riaprire il nodo Plot tempo.
- Rimuovere il campo *Totale* dall'elenco Serie (selezionarlo, quindi fare clic sul pulsante X rosso).
- Aggiungere all'elenco i campi da *Mercato_1* a *Mercato_5*.
- Fare clic su Esegui.

Figura 14-7
Plot tempo di più campi



L'analisi dei singoli mercati evidenzia una tendenza al rialzo stabile in tutti i casi. Alcuni mercati registrano un andamento leggermente più mutevole di altri, ma non vengono evidenziati elementi di stagionalità.

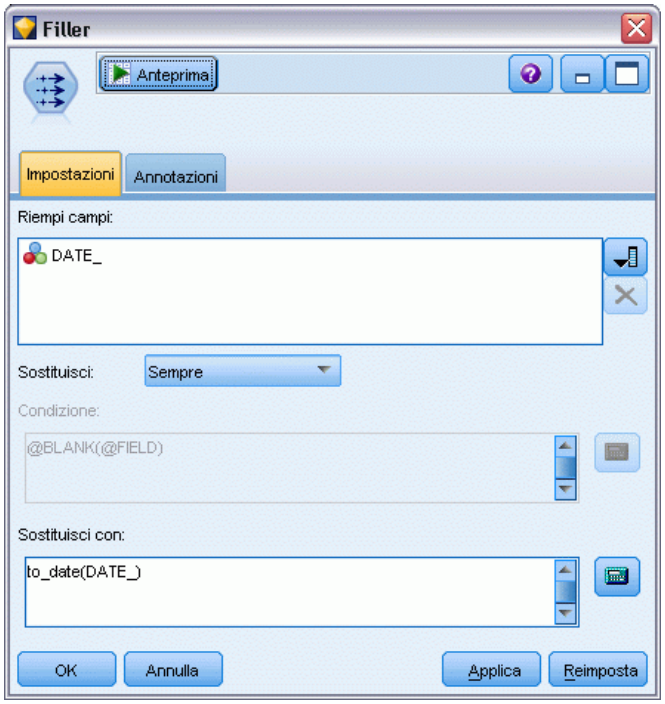
Definizione delle date

A questo punto è necessario modificare il tipo di archiviazione del campo *DATA_* nel formato data.

- Collegare un nodo Riempimento al nodo Filtro.
- Aprire il nodo Riempimento e fare clic sul pulsante di selezione dei campi.
- Selezionare DATA_ per aggiungerlo a Riempi campi.
- Impostare la condizione Sostituisci su Sempre.

▶ Impostare il valore di Sostituisci con su to_date(DATA_).

Figura 14-8
Impostazione del tipo di archiviazione della data

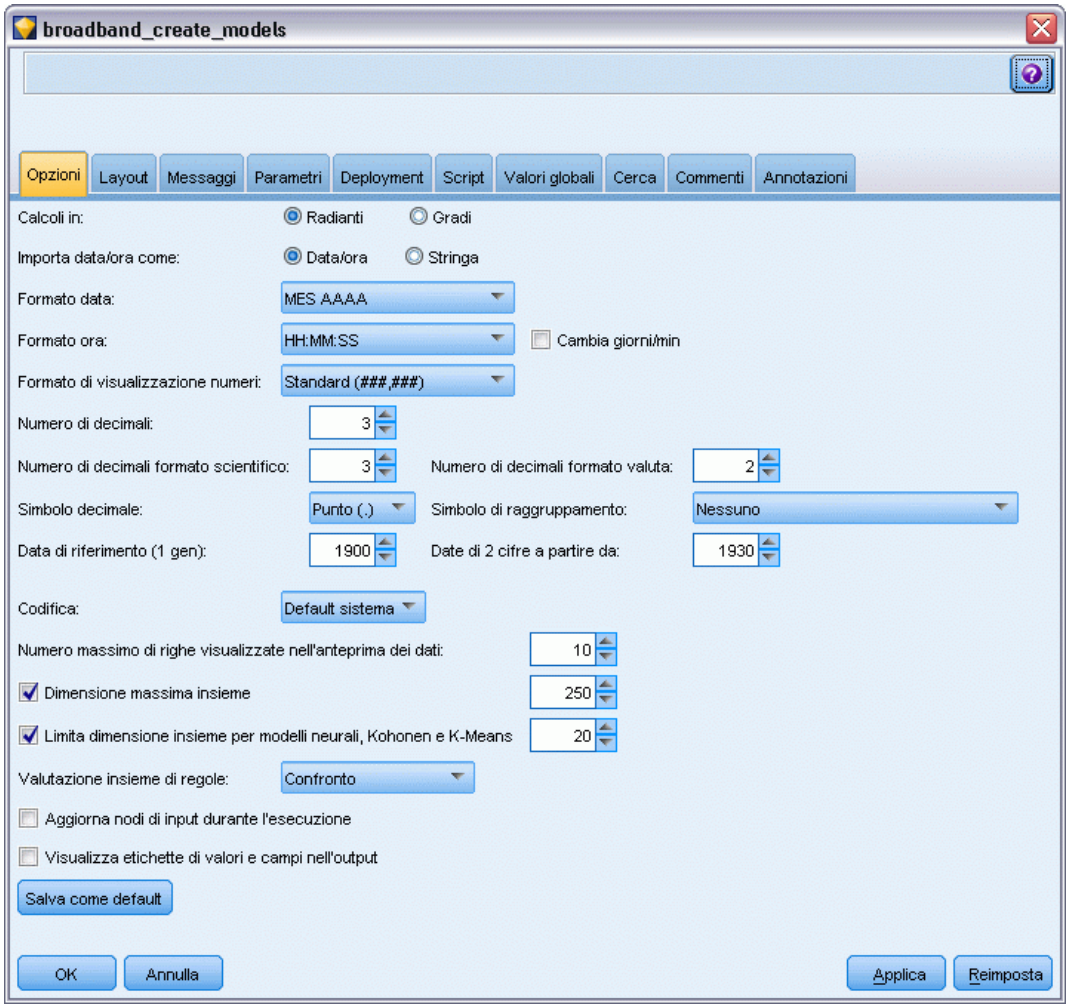


Modificare il formato data di default in modo che corrisponda al formato del campo data. Questa operazione è necessaria perché la conversione del campo data funzioni come previsto.

▶ Nel menu, scegliere Strumenti > Proprietà stream > Opzioni per visualizzare la finestra di dialogo Opzioni stream.

▶ Impostare il Formato data di default su MES AAAA .

Figura 14-9
Impostazione del formato data

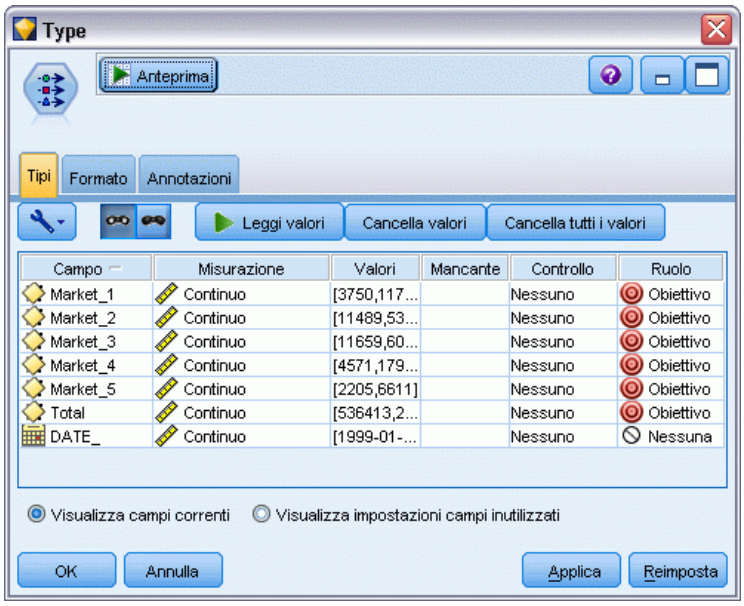


Definizione degli obiettivi

▶ Aggiungere un nodo Tipo e impostare il ruolo del campo *DATA_* su Nessuna. Impostare il ruolo degli altri campi (i campi *Mercato_n* e il campo *Totale*) su Obiettivo.

▶ Fare clic sul pulsante Leggi valori per popolare la colonna Valori.

Figura 14-10
Impostazione del ruolo per più campi



Impostazione degli intervalli di tempo

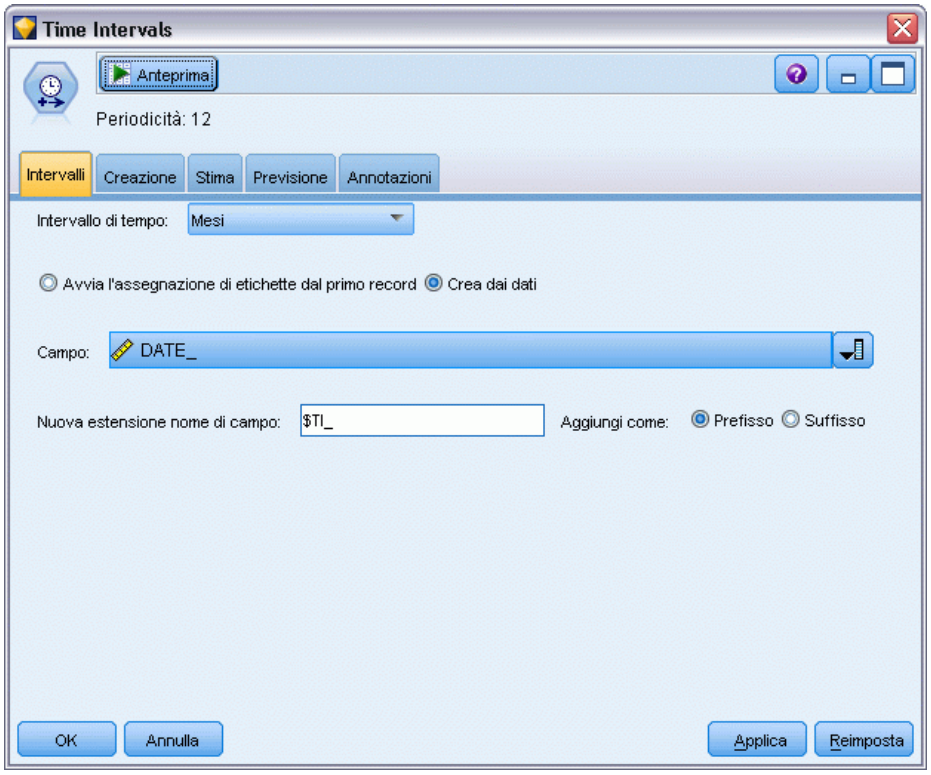
▶ Aggiungere un nodo Intervalli di tempo (dalla palette Operazioni sui campi).

▶ Nella scheda Intervalli, selezionare Mesi come intervallo di tempo.

▶ Selezionare l'opzione Crea dai dati.

- Selezionare DATA_ come campo di creazione.

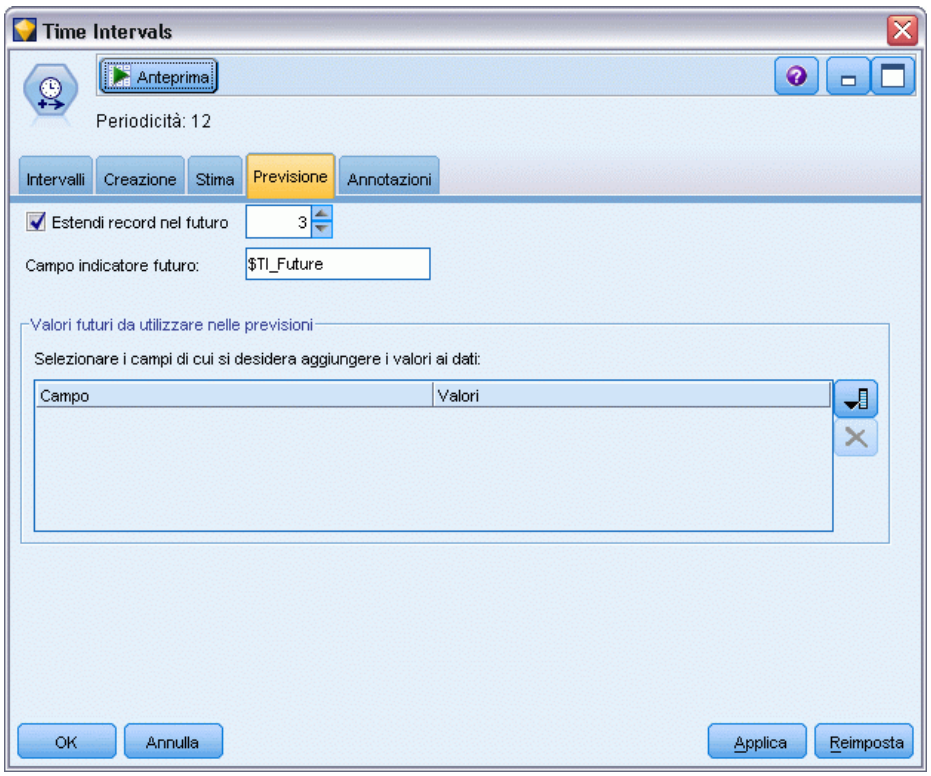
Figura 14-11
Impostazione dell'intervallo di tempo



- Nella scheda Previsione, selezionare la casella di controllo Estendi record nel futuro.
- Impostare il valore su 3.

▶ Fare clic su OK.

Figura 14-12
Impostazione del periodo di previsione

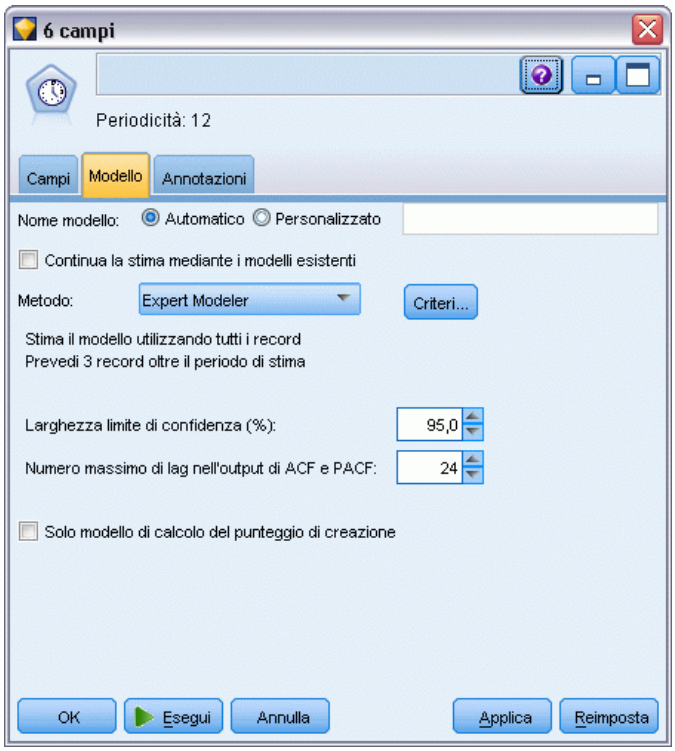


Creazione del modello

▶ Dalla palette Modelli, aggiungere un nodo Serie storica allo stream e collegarlo al nodo Intervalli di tempo.

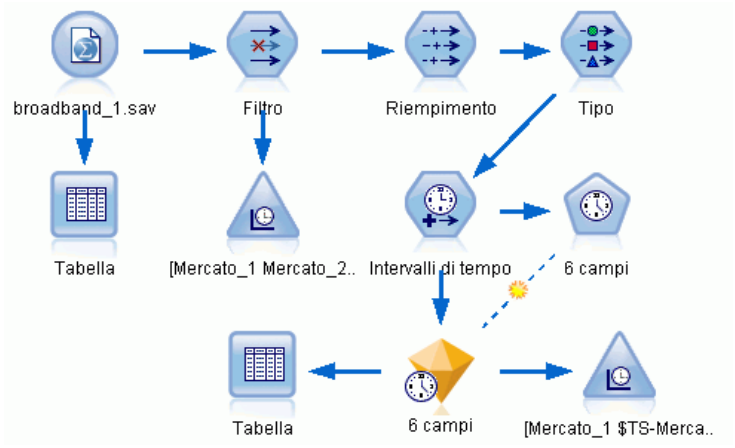
- Fare clic su Esegui nel nodo Serie storica utilizzando tutte le impostazioni di default. In questo modo, Expert Modeler sarà in grado di individuare il modello più appropriato da utilizzare per ogni serie storica.

Figura 14-13
Selezione di Expert Modeler per Serie storica



- Collegare l'insieme di modelli Serie storica al nodo Intervalli di tempo.
- Collegare un nodo Tabella al modello di serie storica e fare clic su Esegui.

Figura 14-14
Stream campione per mostrare la creazione di modelli di serie storica



A questo punto, tre nuove righe (dalla 61 alla 63) sono state accodate ai dati originali. Si tratta delle righe relative al periodo di previsione, in questo caso, da gennaio a marzo 2004.

Sono inoltre presenti diverse nuove colonne: una serie di colonne *$TI_*, aggiunte dal nodo Intervalli di tempo, e le colonne *$TS-*, aggiunte dal nodo Serie storica. Le colonne indicano quanto segue per ogni riga (cioè, ogni intervallo nei dati di serie storica):

Column	**Descrizione**
$TI_IndiceTempo	Valore dell'indice dell'intervallo di tempo per questa riga.
$TI_EtichettaTempo	Etichetta dell'intervallo di tempo per questa riga.
$TI_Anno	Gli indicatori di anno e mese per i dati generati in questa riga.
$TI_Mese	
$TI_Conteggio	Il numero dei record coinvolti nella determinazione dei nuovi dati per questa riga.
$TI_Futuro	Indica se questa riga contiene dati di previsione.
$TS-*nomecolonna*	I dati del modello generato per ogni colonna dei dati originali.
$TSLCI-*nomecolonna*	Il valore di intervallo di confidenza inferiore per ogni colonna dei dati del modello generato.
$TSUCI-*nomecolonna*	Il valore di intervallo di confidenza superiore per ogni colonna dei dati del modello generato.
$TS-Totale	Il totale dei valori della colonna $TS-*nomecolonna* per questa riga.
$TSLCI-Totale	Il totale dei valori della colonna $TSLCI-*nomecolonna* per questa riga.
$TSUCI-Totale	Il totale dei valori della colonna $TSUCI-*nomecolonna* per questa riga.

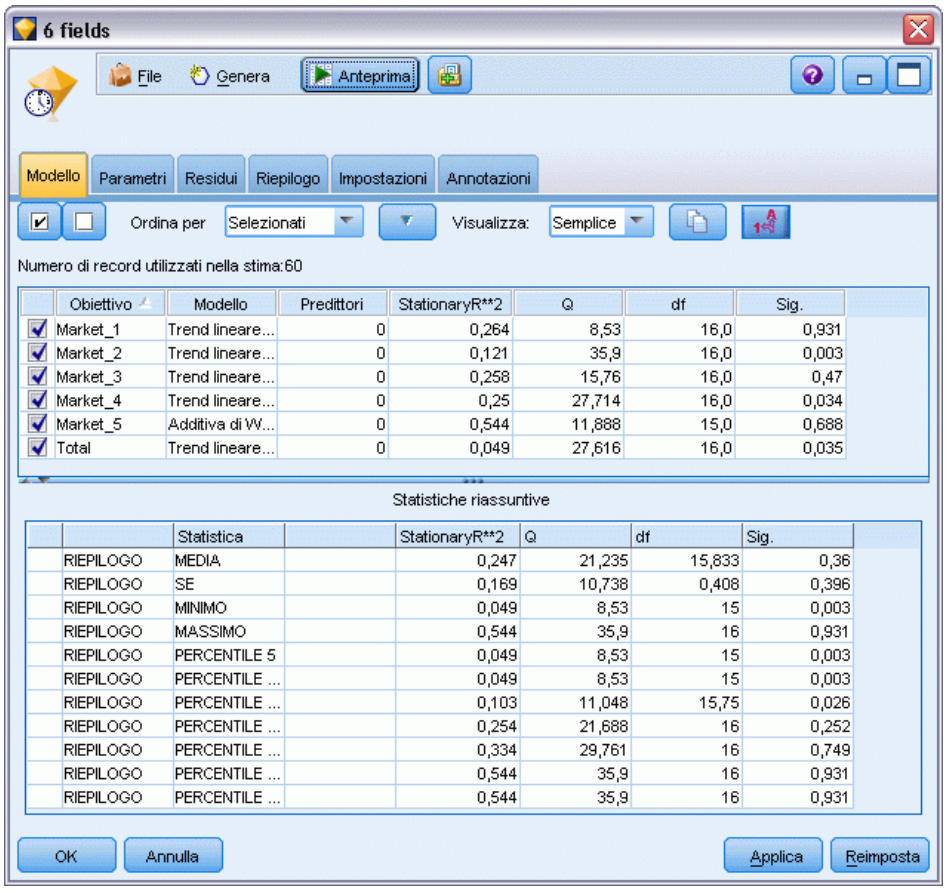
Le colonne più significative per l'operazione di previsione sono le colonne *$TS-Mercato_n*, *$TSLCI-Mercato_n* e *$TSUCI-Mercato_n*. In particolare, nelle righe dalla 61 alla 63, queste colonne contengono i dati di previsione relativi agli abbonamenti degli utenti e gli intervalli di confidenza per ciascuno dei mercati locali.

Esame del modello

- Fare doppio clic sull'insieme di modelli Serie storica per visualizzare i dati relativi ai modelli generati per ciascun mercato.

Si noti che Expert Modeler ha scelto di generare per il Mercato 5 un modello diverso rispetto al tipo di modello generato per gli altri mercati.

Figura 14-15
Modelli di serie storica generati per i mercati



Obiettivo	Modello	Predittori	StationaryR**2	Q	df	Sig.
Market_1	Trend lineare...	0	0,264	8,53	16,0	0,931
Market_2	Trend lineare...	0	0,121	35,9	16,0	0,003
Market_3	Trend lineare...	0	0,258	15,76	16,0	0,47
Market_4	Trend lineare...	0	0,25	27,714	16,0	0,034
Market_5	Additiva di W...	0	0,544	11,888	15,0	0,688
Total	Trend lineare...	0	0,049	27,616	16,0	0,035

	Statistica	StationaryR**2	Q	df	Sig.
RIEPILOGO	MEDIA	0,247	21,235	15,833	0,36
RIEPILOGO	SE	0,169	10,738	0,408	0,396
RIEPILOGO	MINIMO	0,049	8,53	15	0,003
RIEPILOGO	MASSIMO	0,544	35,9	16	0,931
RIEPILOGO	PERCENTILE 5	0,049	8,53	15	0,003
RIEPILOGO	PERCENTILE ...	0,049	8,53	15	0,003
RIEPILOGO	PERCENTILE ...	0,103	11,048	15,75	0,026
RIEPILOGO	PERCENTILE ...	0,254	21,688	16	0,252
RIEPILOGO	PERCENTILE ...	0,334	29,761	16	0,749
RIEPILOGO	PERCENTILE ...	0,544	35,9	16	0,931
RIEPILOGO	PERCENTILE ...	0,544	35,9	16	0,931

La colonna Predittori mostra il numero dei campi utilizzati come predittori per i singoli obiettivi—in questo caso, nessuno.

Le restanti colonne in questa visualizzazione mostrano diverse misure di bontà di adattamento per ogni modello. La colonna StationaryR**2 mostra il valore *R*-quadrato stazionaria. Questa statistica offre una stima della proporzione della variazione totale nella serie che viene spiegata nel modello. A un valore superiore (fino a un massimo di 1.0) corrisponde un migliore adattamento del modello.

Le colonne Q, df e Sig. si riferiscono alla statistica Ljung-Box, un test della casualità degli errori residui nel modello. Maggiore è la casualità gli errori, più probabile è che il modello sia migliore. Q è la stessa statistica Ljung-Box, mentre df (degrees of freedom, gradi di libertà) indica il numero di parametri del modello che sono liberi di variare quando si stima un determinato obiettivo.

La colonna Sig. fornisce il valore di significatività della statistica Ljung-Box, che rappresenta un'ulteriore indicazione della correttezza della specificazione del modello. Un valore di significatività inferiore a 0.05 implica che gli errori residui non sono casuali e che nella serie osservata è presente una struttura che non viene spiegata dal modello.

Tenendo conto di entrambi i valori *R*-quadrato stazionaria e Sig., i modelli scelti da Expert Modeler per *Mercato_1*, *Mercato_3* e *Mercato_5* sono sufficientemente accettabili. I valori Sig. relativi a *Mercato_2* e *Mercato_4* sono entrambi inferiori a 0.05, ovvero per questi mercati potrebbe essere necessario sperimentare modelli con un adattamento migliore.

I valori di riepilogo nella parte inferiore della visualizzazione forniscono informazioni sulla distribuzione delle statistiche di tutti i modelli. Per esempio, il valore medio di *R*-quadrato stazionaria di tutti i modelli è pari a 0.247, mentre il valore minimo è 0.049 (relativo al modello *Totale*) e il valore massimo è 0.544 (relativo a *Mercato_5*).

SE indica l'errore standard calcolato su tutti i modelli per ogni statistica. Per esempio, l'errore standard per *R*-quadrato stazionaria di tutti i modelli è pari a 0.169.

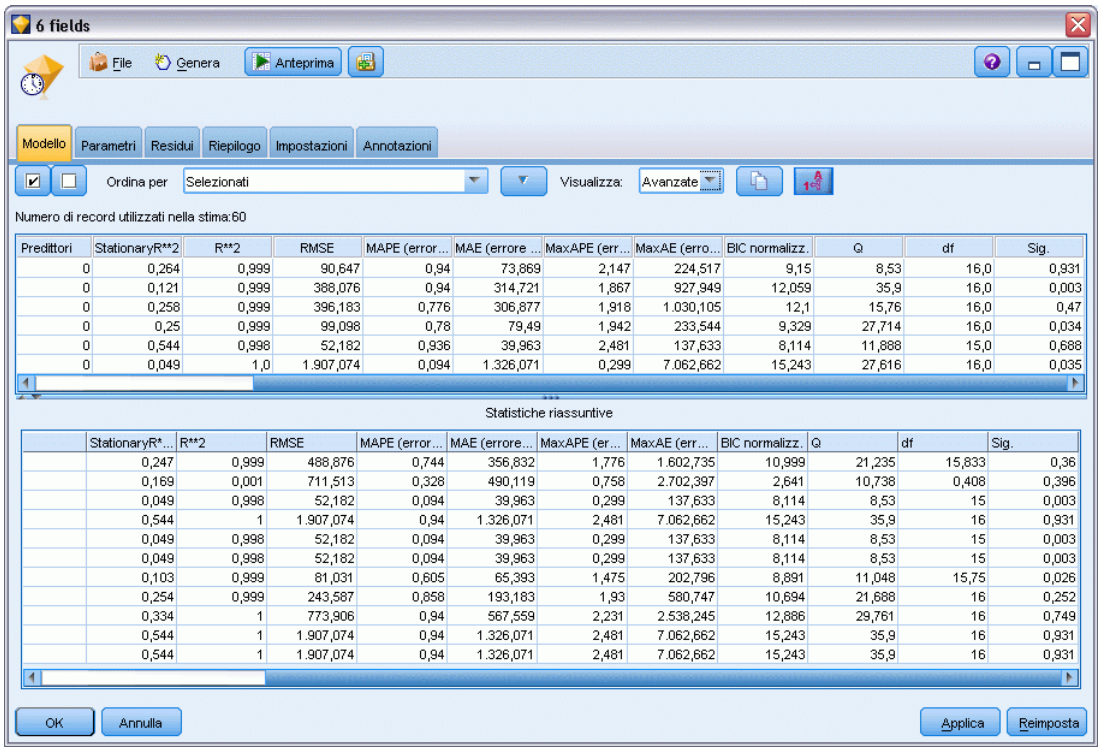
La sezione di riepilogo contiene anche i percentili che forniscono informazioni sulla distribuzione delle statistiche dei modelli. Per ciascun percentile, una certa percentuale di modelli ha un valore statistico di adattamento inferiore al valore specificato.

Quindi, per esempio, solo il 25% dei modelli presenta un valore *R*-quadrato stazionaria minore di 0.121.

▶ Fare clic sull'elenco a discesa Visualizza e selezionare Opzioni avanzate.

Vengono visualizzate svariate misure della bontà di adattamento aggiuntive. R**2 è il valore *R*-quadrato, una stima della variazione totale nella serie storica che viene spiegata nel modello. Poiché il valore massimo per questa statistica è 1.0, i modelli vanno bene sotto questo aspetto.

Figura 14-16
Visualizzazione delle opzioni avanzate dei modelli di serie storica



RMSE è la radice errore quadratico medio, una misura di quanto i valori effettivi di una serie differiscono dai valori previsti dal modello ed è espresso nelle stesse unità di quelle utilizzate per la serie stessa. Poiché è la misura di un errore, il suo valore dovrebbe essere il più basso possibile. A prima vista sembra che i modelli per *Mercato_2* e *Mercato_3*, sebbene ancora accettabili in base alle statistiche considerate fino ad ora, siano meno efficaci di quelli per gli altri tre mercati.

Queste misure aggiuntive della bontà di adattamento includono gli errori assoluti percentuali medio (MAPE) e massimo (MaxAPE). L'errore assoluto percentuale rappresenta una misura del grado di variazione di una serie obiettivo rispetto al livello previsto dal relativo modello, espresso in valore percentuale. L'analisi degli errori medi e massimi dei modelli permette di avere un'indicazione del grado di incertezza delle previsioni.

Il valore MAPE (errore assoluto medio percentuale) indica che tutti i modelli presentano un'incertezza media inferiore all'1%, che è un valore molto basso. Il valore MaxAPE indica l'errore assoluto percentuale massimo ed è utile per immaginare lo scenario peggiore relativo alle previsioni. Tale valore mostra che l'errore percentuale massimo per ciascuno dei modelli è più o meno compreso nell'intervallo 1.8 - 2.5%, anch'esso un insieme di valori molto basso.

Il valore MAE (errore assoluto medio) mostra la media dei valori assoluti degli errori di previsione. Analogamente al valore RMSE, è espresso nelle stesse unità di quelle utilizzate per la serie stessa. Il valore MaxAE mostra l'errore di previsione più grande espresso nelle stesse unità e indica lo scenario peggiore per le previsioni.

Per quanto interessanti possano essere questi valori assoluti, quelli più utili in questo caso sono i valori degli errori percentuali (MAPE e MaxAPE), poiché le serie obiettivo rappresentano i numeri di abbonati per i mercati di varie dimensioni.

I valori MAPE e MaxAPE indicano un grado di incertezza accettabile? Si tratta senza dubbio di valori molto bassi. Una situazione di questo tipo richiede comunque una valutazione aziendale, in quanto il grado di accettabilità del rischio varia a seconda dei diversi problemi. Si presumerà che le statistiche della bontà di adattamento rientrino entro limiti accettabili e si procederà all'esame gli errori residui.

Una disamina delle autocorrelazioni (ACF) e autocorrelazioni parziali (PACF) dei residui dei modelli offre una visione dei modelli quantitativamente più attendibile rispetto alla semplice visualizzazione delle statistiche relative alla bontà di adattamento.

Un modello Serie temporale ben specificato catturerà tutte le variazioni non casuali, incluse stagionalità, trend, ciclicità e altri fattori importanti. In questo caso, un eventuale errore non deve essere correlato a se stesso (autocorrelato) nel tempo. Una struttura significativa in una delle due funzioni di autocorrelazione implicherebbe che il modello sottostante è incompleto.

▶ Fare clic sulla scheda Residui per visualizzare i valori delle funzioni di autocorrelazione (ACF) e autocorrelazione parziale (PACF) per gli errori residui nel modello per il primo dei mercati locali.

Figura 14-17
Valori ACF e PACF per i mercati

In questi grafici plot, i valori originali della variabile errore sono stati ritardati fino a 24 periodi di tempo e confrontati con il valore originale per vedere se esiste un'eventuale correlazione nel tempo. Affinché il modello sia accettabile, nessuna delle barre nel grafico plot superiore (ACF) deve estendersi oltre l'area ombreggiata, in direzione positiva (verso l'alto) o negativa (verso il basso).

In questo caso, sarebbe necessario controllare il grafico plot inferiore (PACF) per vedere se la struttura è confermata. Il grafico plot PACF osserva le correlazioni dopo aver controllato i valori della serie ai relativi punti temporali.

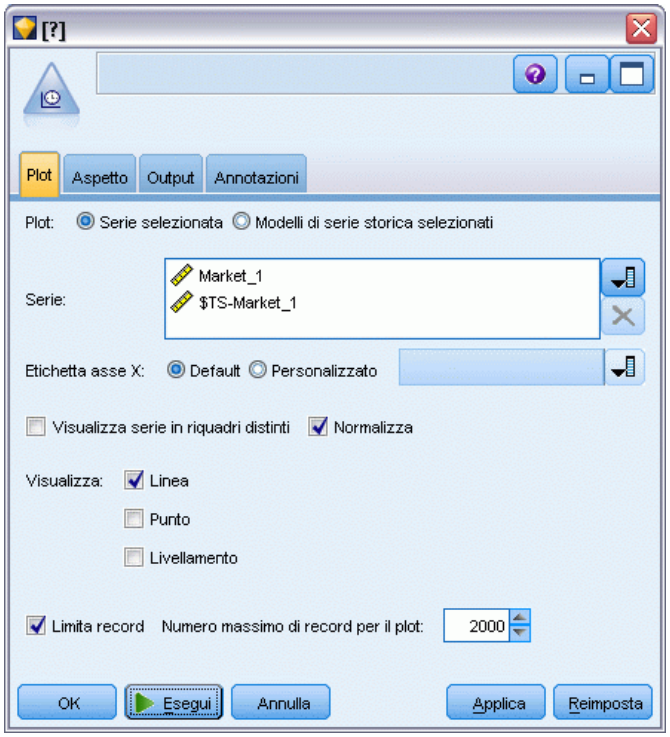
I valori per *Mercato_1* ricadono tutti nell'area ombreggiata. È dunque possibile proseguire controllando i valori per gli altri mercati.

▶ Fare clic sull'elenco a discesa Visualizza plot del modello per visualizzare questi valori per gli altri mercati e i totali.

I valori per *Mercato_2* e *Mercato_4* offrono motivi di preoccupazione confermando ciò che si era sospettato in precedenza per i rispettivi valori Sig.. A un certo punto per questi mercati si dovranno sperimentare altri modelli per verificare se è possibile trovare un adattamento migliore, tuttavia nel prosieguo di questo esempio ci si concentrerà su quanto è possibile ancora apprendere dal modello *Mercato_1*.

- Dalla palette Grafici, collegare un nodo Plot tempo all'insieme di modelli Serie storica.
- Nella scheda Plot, deselezionare la casella di controllo Visualizza serie in riquadri distinti.
- Nell'elenco Serie, fare clic sul pulsante di selezione dei campi, selezionare i campi *Mercato_1* e *$TS-Mercato_1* e fare clic su OK per aggiungerli all'elenco.
- Fare clic su Esegui per visualizzare un grafico a linee dei dati effettivi e previsti relativi al primo mercato locale.

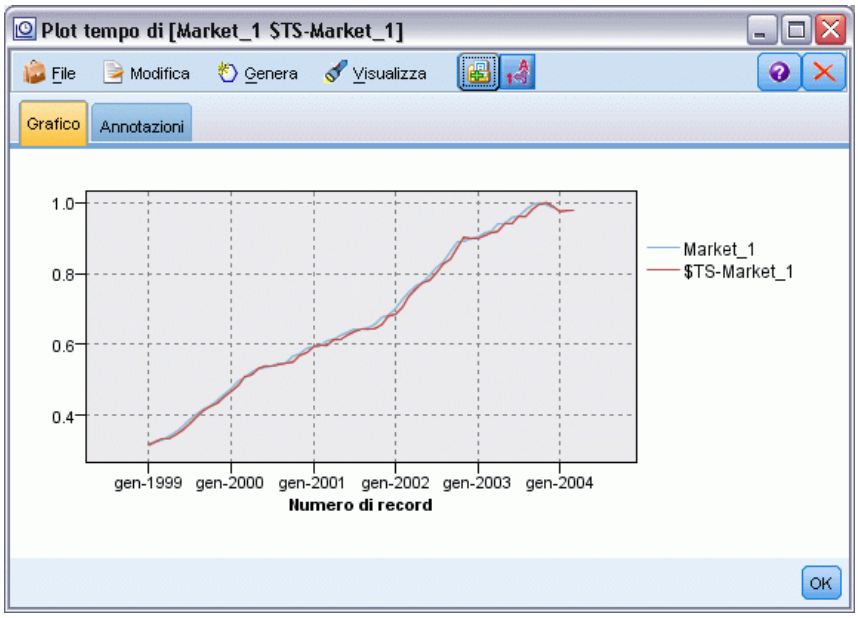
Figura 14-18
Selezione dei campi da rappresentare nel grafico



Si noti come la linea della previsione (*$TS-Mercato_1*) si estende oltre il punto in cui terminano i dati effettivi. A questo punto, si dispone di una stima della domanda di mercato prevista per i prossimi tre mesi in questo mercato.

Le linee del grafico relative ai dati effettivi e previsti su tutta la serie storica sono molto vicine, a indicare che il modello è affidabile per questa serie storica specifica.

Figura 14-19
Plot tempo dei dati effettivi e previsti per Mercato_1



Salvare il modello in un file per poterlo riutilizzare in un esempio futuro:

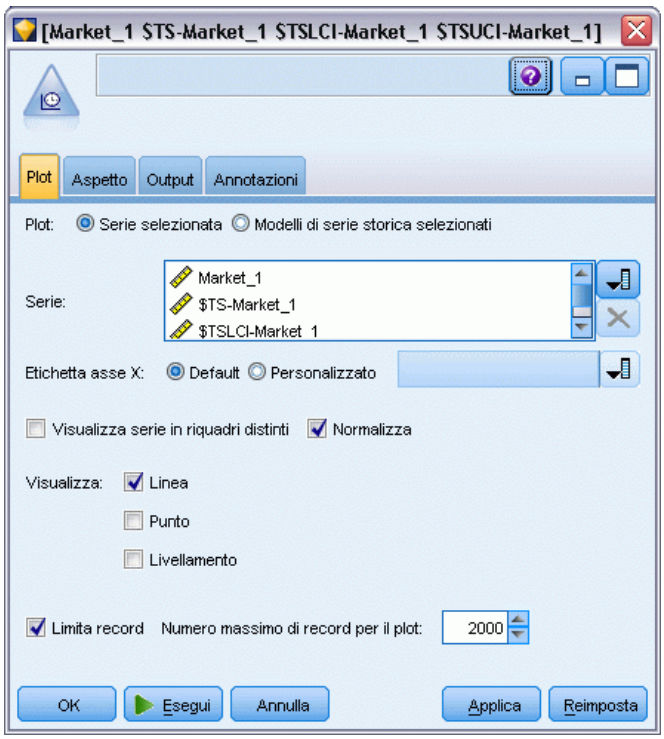
- Fare clic su OK per chiudere il grafico corrente.
- Aprire l'insieme di modelli Serie storica.
- Scegliere File > Salva nodo e specificare il percorso del file.
- Scegliere Salva.

Il modello ottenuto per questo mercato specifico è affidabile, ma qual è il margine d'errore della previsione? Per avere un'indicazione in questo senso, è possibile esaminare l'intervallo di confidenza.

- Fare doppio clic sul nodo Plot tempo nello stream (quello denominato Mercato_1 $TS-Mercato_1) per aprire nuovamente la relativa finestra di dialogo.
- Fare clic sul pulsante di selezione dei campi e aggiungere i campi *$TSLCI-Mercato_1* e *$TSUCI-Mercato_1* all'elenco Serie.

▶ Fare clic su Esegui.

Figura 14-20
Aggiunta di altri campi da rappresentare nel grafico

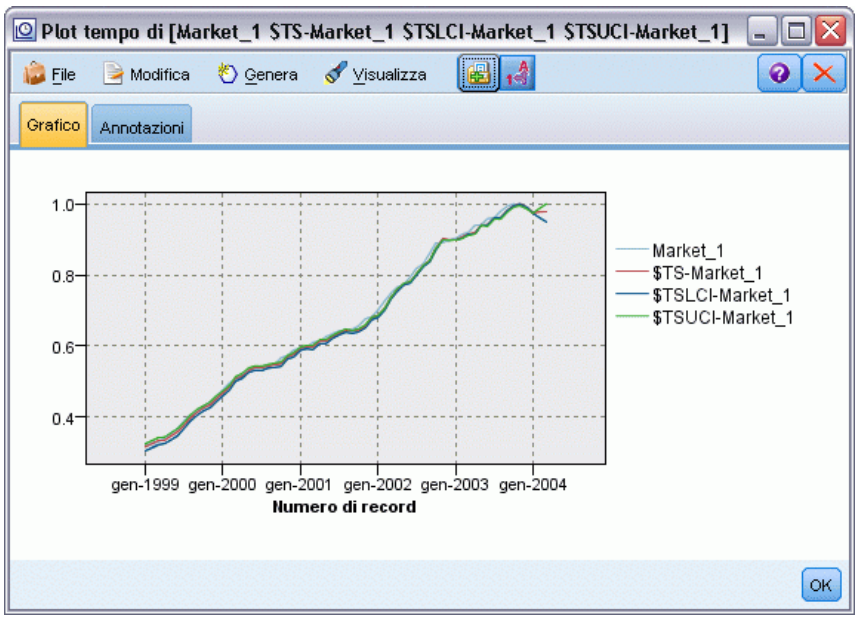


A questo punto si dispone dello stesso grafico di prima, al quale sono stati aggiunti i limiti superiore (*$TSUCI*) e inferiore (*$TSLCI*) dell'intervallo di confidenza.

Si noti come i limiti dell'intervallo di confidenza divergono nel corso del periodo di previsione, a indicare il crescente grado di incertezza con il protrarsi delle previsioni nel futuro.

Tuttavia, con il passare di ogni periodo di tempo, si disporrà (in questo caso) di un altro mese di dati effettivi di utilizzo sui quali basare le previsioni. È possibile leggere i nuovi dati in uno stream e riapplicare il modello, dal momento che si è capito che si tratta di un modello affidabile. Per ulteriori informazioni, vedere l'argomento Riapplicazione di un modello di serie storica a pag. 197.

Figura 14-21
Plot tempo con intervallo di confidenza aggiunto



Riepilogo

Finora, è stato spiegato come utilizzare Expert Modeler per produrre previsioni relative a più serie storiche e salvare i modelli risultanti in un file esterno.

Nell'esempio successivo, viene spiegato come trasformare i dati delle serie storiche non standard in un formato adeguato all'input in un nodo Serie storica.

Riapplicazione di un modello di serie storica

Questo esempio consente di applicare i modelli di serie storica del primo esempio ma può anche essere utilizzato in modo indipendente. Per ulteriori informazioni, vedere l'argomento Previsione mediante il nodo Serie storica a pag. 175.

Come nel primo scenario, un analista di una società che fornisce accesso a banda larga su scala nazionale deve generare previsioni mensili degli abbonamenti nei singoli mercati locali al fine di prevedere l'utilizzo della larghezza di banda. Si supponga di aver già usato Expert Modeler per creare modelli e previsioni per i successivi tre mesi,

Il data warehouse è stato aggiornato con i dati effettivi relativi al periodo di previsione originario; quindi, si desidera utilizzare questi dati per estendere l'orizzonte di previsione di altri tre mesi.

In questo esempio viene utilizzato lo stream denominato *broadband_apply_model.str*, che fa riferimento al file di dati *broadband2.sav*. Questi file sono disponibili nella cartella *Demos* di tutte le installazioni di IBM® SPSS® Modeler. L'accesso si effettua dal gruppo di programmi IBM® SPSS® Modeler nel menu Start di Windows. Il file *broadband_apply_models.str* si trova nella cartella *streams*.

Recupero dello stream

In questo esempio, viene ricreato un nodo Serie storica dal modello di serie storica salvato nel primo esempio. Se non si dispone di un modello salvato è possibile comunque procedere utilizzando il modello fornito nella cartella *Demos*.

- Aprire lo stream *broadband_apply_models.str* nella cartella *streams* all'interno di *Demos*.

Figura 14-22
Apertura dello stream

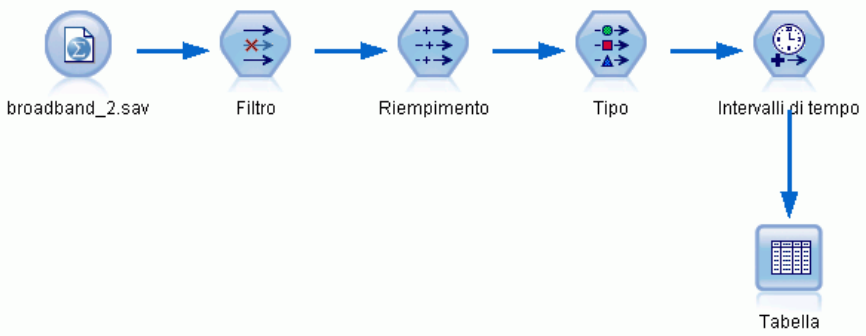
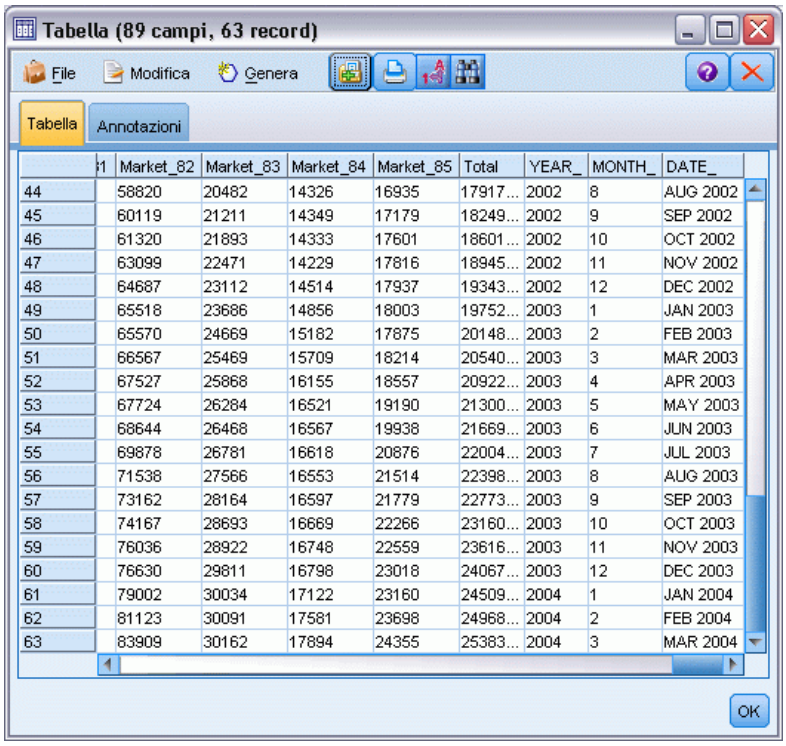


Figura 14-23
Dati di vendita aggiornati



	Market_82	Market_83	Market_84	Market_85	Total	YEAR_	MONTH_	DATE_
44	58820	20482	14326	16935	17917...	2002	8	AUG 2002
45	60119	21211	14349	17179	18249...	2002	9	SEP 2002
46	61320	21893	14333	17601	18601...	2002	10	OCT 2002
47	63099	22471	14229	17816	18945...	2002	11	NOV 2002
48	64687	23112	14514	17937	19343...	2002	12	DEC 2002
49	65518	23686	14856	18003	19752...	2003	1	JAN 2003
50	65570	24669	15182	17875	20148...	2003	2	FEB 2003
51	66567	25469	15709	18214	20540...	2003	3	MAR 2003
52	67527	25868	16155	18557	20922...	2003	4	APR 2003
53	67724	26284	16521	19190	21300...	2003	5	MAY 2003
54	68644	26468	16567	19938	21669...	2003	6	JUN 2003
55	69878	26781	16618	20876	22004...	2003	7	JUL 2003
56	71538	27566	16553	21514	22398...	2003	8	AUG 2003
57	73162	28164	16597	21779	22773...	2003	9	SEP 2003
58	74167	28693	16669	22266	23160...	2003	10	OCT 2003
59	76036	28922	16748	22559	23616...	2003	11	NOV 2003
60	76630	29811	16798	23018	24067...	2003	12	DEC 2003
61	79002	30034	17122	23160	24509...	2004	1	JAN 2004
62	81123	30091	17581	23698	24968...	2004	2	FEB 2004
63	83909	30162	17894	24355	25383...	2004	3	MAR 2004

I dati mensili aggiornati vengono raccolti in *broadband_2.sav*.

- Collegare un nodo Tabella al nodo File di input IBM® SPSS® Statistics, aprire il nodo Tabella e fare clic su Esegui.

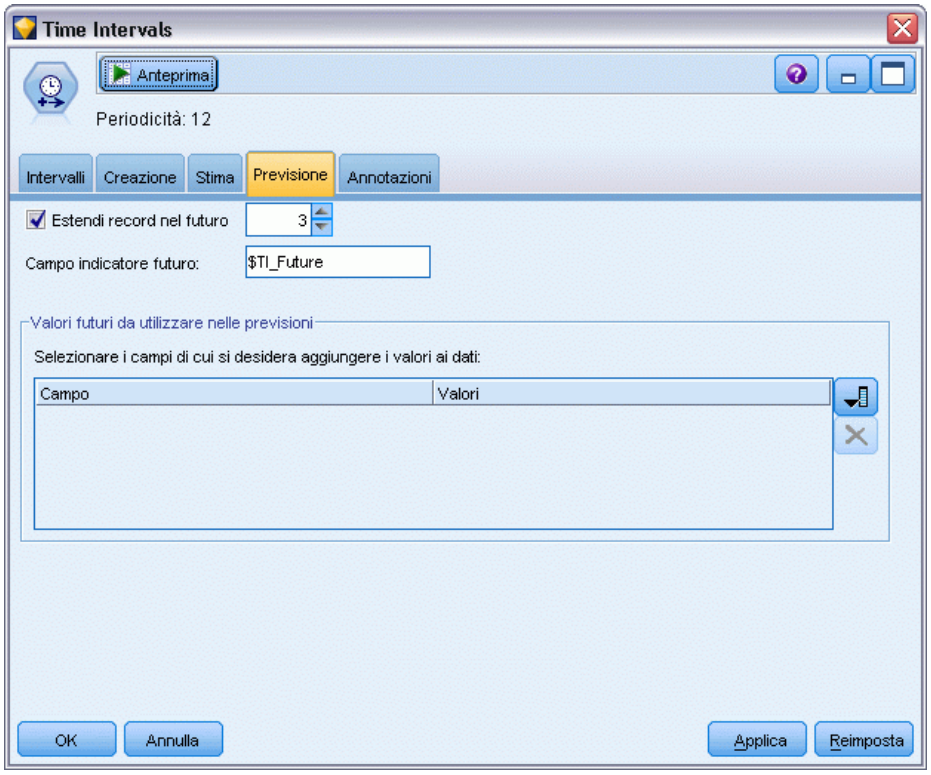
  Nota: il file di dati è stato aggiornato con i dati effettivi delle vendite relative al periodo di gennaio, febbraio e marzo 2004, nelle righe 61, 62, 63.

- Aprire il nodo Intervalli di tempo nello stream.

- Fare clic sulla scheda Previsione.

▶ Verificare che Estendi record nel futuro sia impostato su 3.

Figura 14-24
Verifica dell'impostazione del periodo di previsione

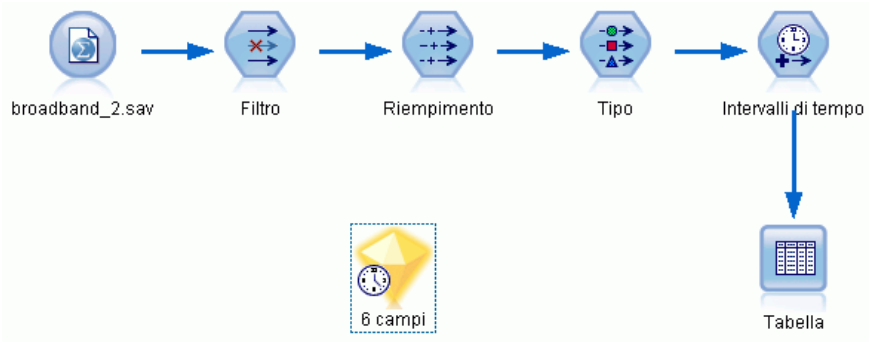


Recupero del modello salvato

▶ Dal menu IBM® SPSS® Modeler, scegliere Inserisci > Nodo dal file, quindi selezionare il file *TSmodel.nod* dalla cartella *Demos* (oppure utilizzare il modello Serie storica salvato nel primo esempio).

Questo file contiene i modelli di serie storica dell'esempio precedente. L'operazione di inserimento posiziona l'insieme di modelli Serie storica corrispondente nell'area di disegno.

Figura 14-25
Aggiunta dell'insieme di modelli

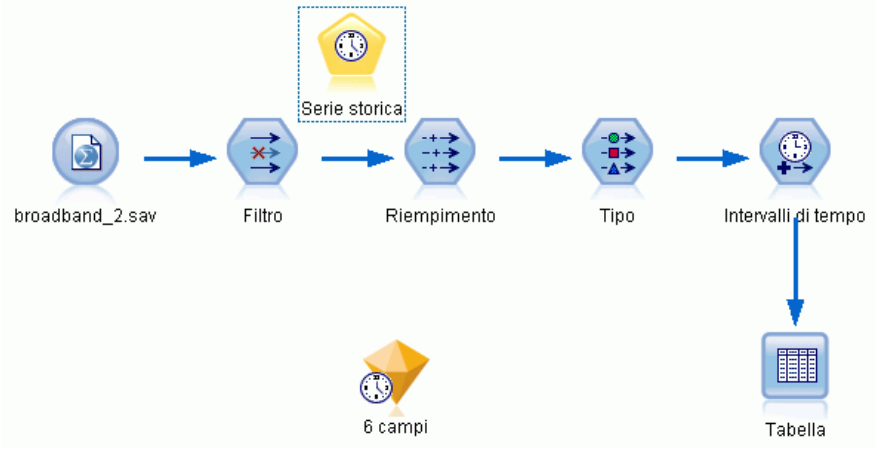


Generazione di un nodo Modelli

- Aprire l'insieme di modelli Serie storica e scegliere Genera > Genera nodo Modelli.

Questa operazione consente di posizionare un nodo Modelli Serie storica nell'area di disegno.

Figura 14-26
Creazione di un nodo Modelli dall'insieme di modelli



Generazione di un nuovo modello

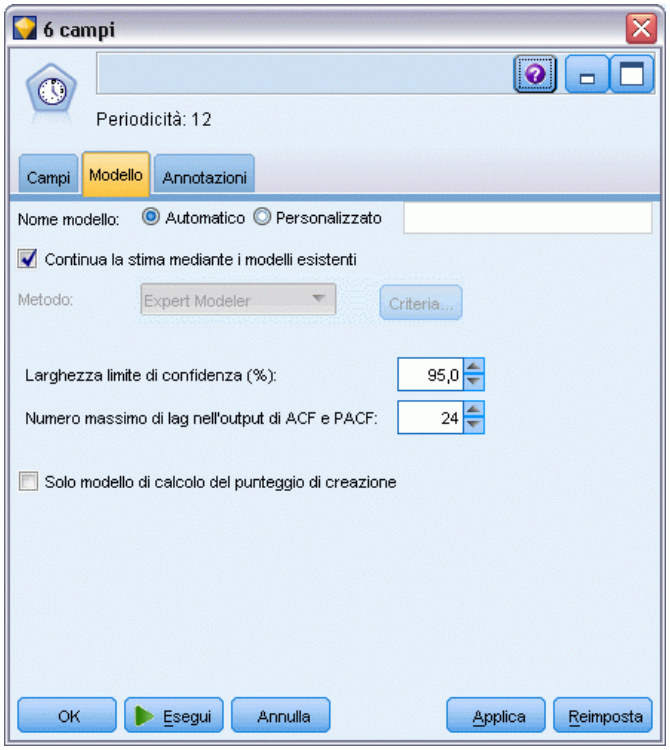
- Chiudere l'insieme di modelli Serie storica ed eliminarlo dall'area di disegno.

  Il modello precedente era stato creato sulla base di 60 righe di dati. È necessario ora creare un nuovo modello in base ai dati di vendita aggiornati (63 righe).

- Collegare allo stream il nodo di creazione di serie storica appena generato.

Figura 14-27
Collegamento del nodo Modelli allo stream

Figura 14-28
Riutilizzo delle impostazioni archiviate per il modello di serie storica



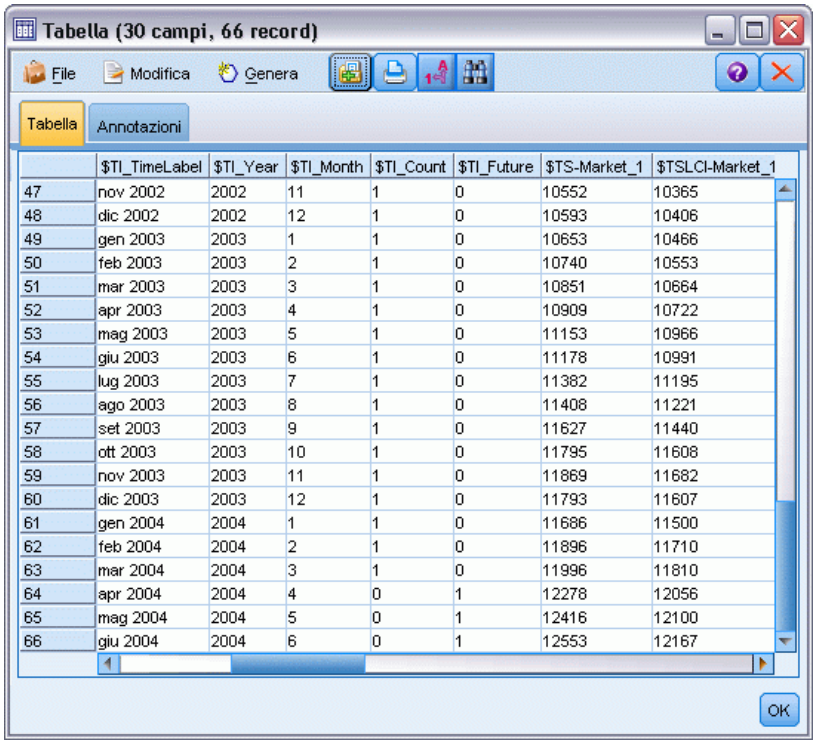
- Aprire il nodo Serie storica.

- Nella scheda Modello, assicurarsi che sia selezionata l'opzione Continua la stima mediante i modelli esistenti.

- Fare clic su Esegui per posizionare un nuovo insieme di modelli nell'area di disegno e nella palette Modelli.

Esame del nuovo modello

Figura 14-29
Tabella contenente le nuove previsioni

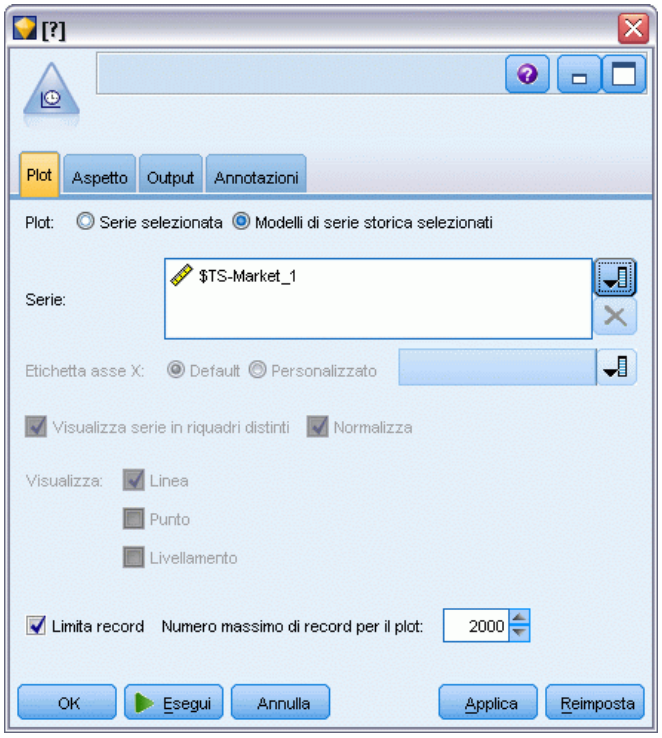


	$TI_TimeLabel	$TI_Year	$TI_Month	$TI_Count	$TI_Future	$TS-Market_1	$TSLCI-Market_1
47	nov 2002	2002	11	1	0	10552	10365
48	dic 2002	2002	12	1	0	10593	10406
49	gen 2003	2003	1	1	0	10653	10466
50	feb 2003	2003	2	1	0	10740	10553
51	mar 2003	2003	3	1	0	10851	10664
52	apr 2003	2003	4	1	0	10909	10722
53	mag 2003	2003	5	1	0	11153	10966
54	giu 2003	2003	6	1	0	11178	10991
55	lug 2003	2003	7	1	0	11382	11195
56	ago 2003	2003	8	1	0	11408	11221
57	set 2003	2003	9	1	0	11627	11440
58	ott 2003	2003	10	1	0	11795	11608
59	nov 2003	2003	11	1	0	11869	11682
60	dic 2003	2003	12	1	0	11793	11607
61	gen 2004	2004	1	1	0	11686	11500
62	feb 2004	2004	2	1	0	11896	11710
63	mar 2004	2004	3	1	0	11996	11810
64	apr 2004	2004	4	0	1	12278	12056
65	mag 2004	2004	5	0	1	12416	12100
66	giu 2004	2004	6	0	1	12553	12167

- Collegare un nodo Tabella al nuovo insieme di modelli di serie storica nell'area di disegno.
- Aprire il nodo Tabella e fare clic su Esegui.

  Il nuovo modello effettua ancora le previsioni su tre mesi in quanto si stanno riutilizzando le impostazioni archiviate. Tuttavia, questa volta, le previsioni si riferiscono al periodo da aprile a giugno in quanto il periodo di stima (specificato nel nodo Intervalli di tempo) ora termina a marzo invece che a gennaio.

Figura 14-30
Specifica dei campi da rappresentare nel grafico



- Collegare un nodo grafico Plot tempo all'insieme di modelli Serie storica.

  Questa volta verrà utilizzata la visualizzazione plot tempo progettata appositamente per i modelli di serie storica.

- Nella scheda Plot, scegliere l'opzione Modelli di serie storica selezionati.

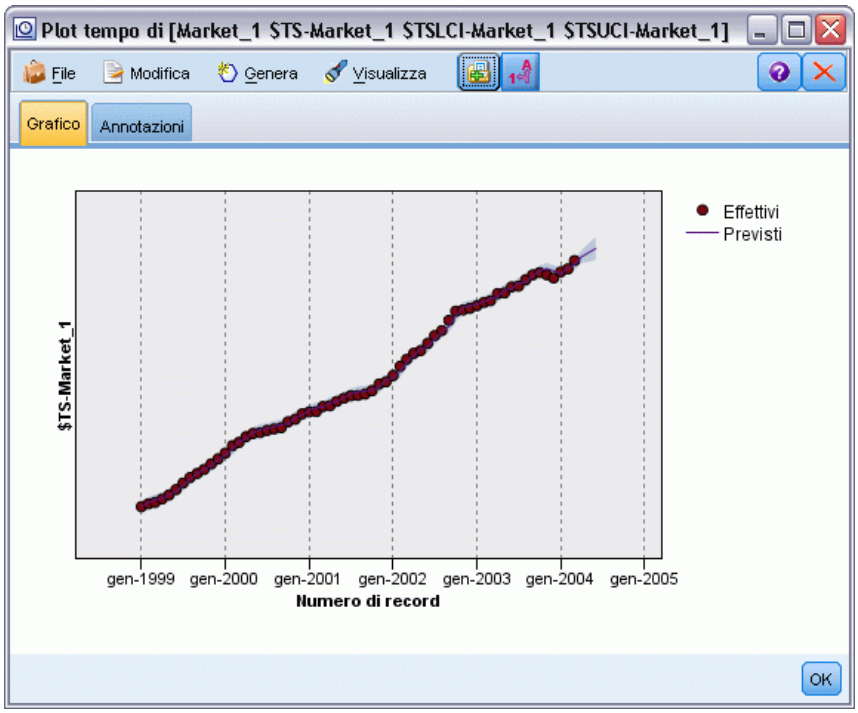
- Nell'elenco Serie, fare clic sul pulsante di selezione dei campi, selezionare il campo *$TS-Mercato_1* e fare clic su OK per aggiungerlo all'elenco.

- Fare clic su Esegui.

  A questo punto, si dispone di un grafico nel quale sono indicate le vendite effettive relative al *Mercato_1* fino al mese di marzo 2004, assieme alle vendite previste e all'intervallo di confidenza (indicato dall'area con ombreggiatura blu) fino al mese di giugno 2004.

Come nel primo esempio, i valori di previsione sono molto simili ai dati effettivi nell'arco del periodo, il che significa, ancora, che si tratta di un modello affidabile.

Figura 14-31
Previsioni estese al mese di giugno



Riepilogo

È stato spiegato come applicare i modelli salvati per estendere le previsioni precedenti nel momento in cui diventano disponibili dei dati più aggiornati, senza dover creare nuovamente i modelli. Ovviamente, se ci sono motivi per credere che un modello sia cambiato, sarà necessario ricrearlo.

Capitolo
15

Previsione delle vendite da catalogo (Serie storica)

Una società di vendite mediante catalogo intende effettuare una previsione delle vendite mensili della linea di abbigliamento per uomo, sulla base dei dati di vendita degli ultimi 10 anni.

In questo esempio viene utilizzato lo stream denominato *catalog_forecast.str*, che fa riferimento al file di dati *catalog_seasfac.sav*. Questi file sono disponibili nella directory *Demos* di tutte le installazioni di IBM® SPSS® Modeler. L'accesso si effettua dal gruppo di programmi IBM® SPSS® Modeler nel menu Start di Windows. Il file *catalog_forecast.str* si trova nella directory *streams*.

In un esempio precedente si è visto che è possibile lasciare a Expert Modeler la scelta del modello più appropriato per la propria serie storica. In questo esempio si esamineranno i due metodi disponibili per la scelta personale del modello: livellamento esponenziale e ARIMA.

Prima di decidere quale sia il modello più appropriato, è utile rappresentare graficamente la serie storica. L'ispezione visiva di una serie storica può facilitare notevolmente la decisione. In modo particolare, è opportuno porsi le seguenti domande:

- La serie presenta un trend generale? In caso affermativo, il trend è costante o sembra attenuarsi con l'andare del tempo?
- La serie presenta caratteristiche di stagionalità? In caso affermativo, le fluttuazioni stagionali aumentano con il passare del tempo oppure appaiono costanti nei vari periodi?

Creazione dello stream

▶ Creare un nuovo stream e aggiungere un nodo di input File Statistics che punti al file *catalog_seasfac.sav*.

Figura 15-1
Previsione delle vendite da catalogo

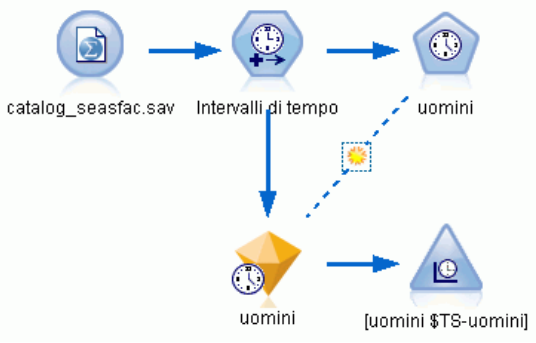
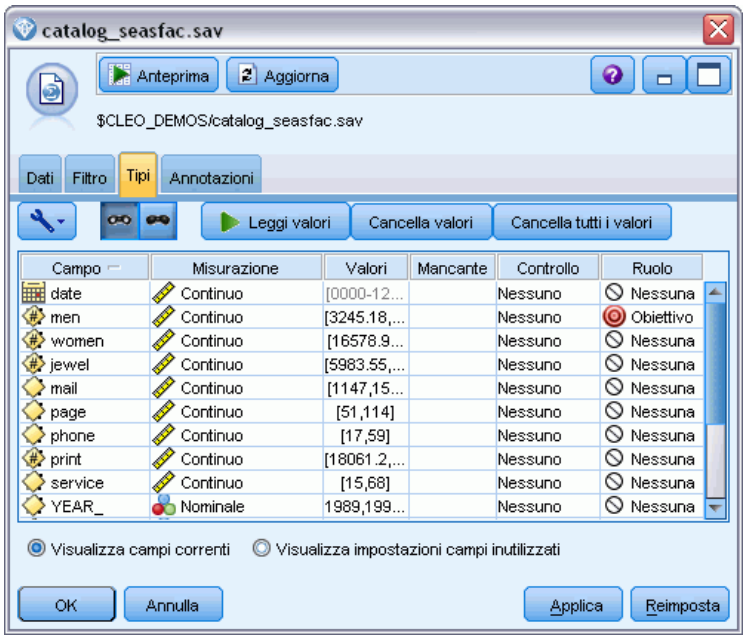
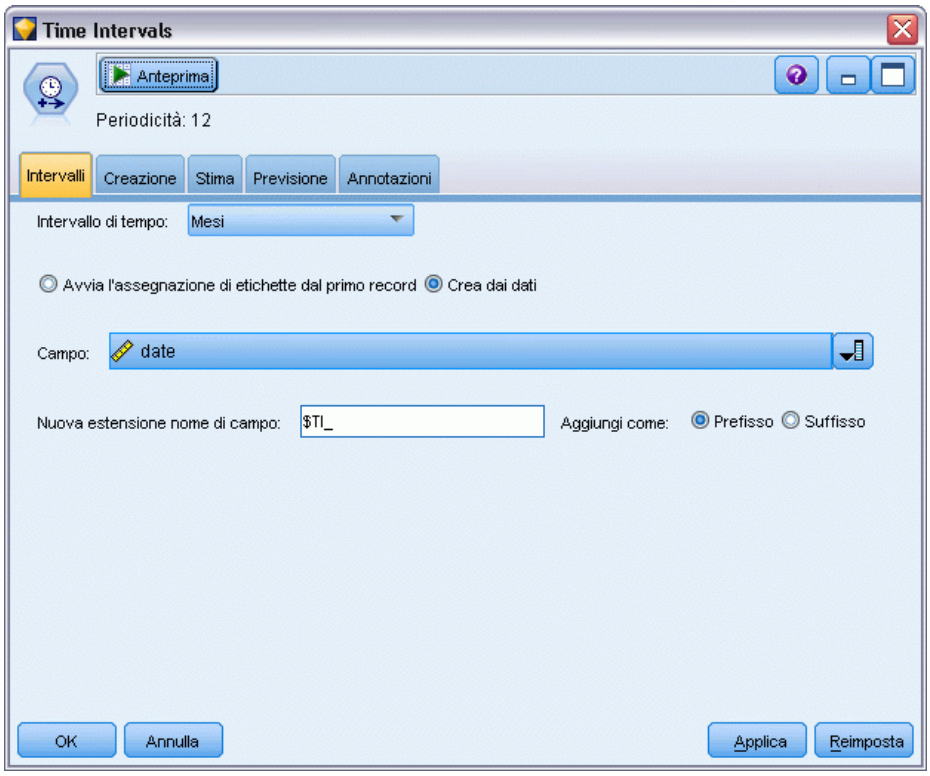


Figura 15-2
Specifica del campo obiettivo



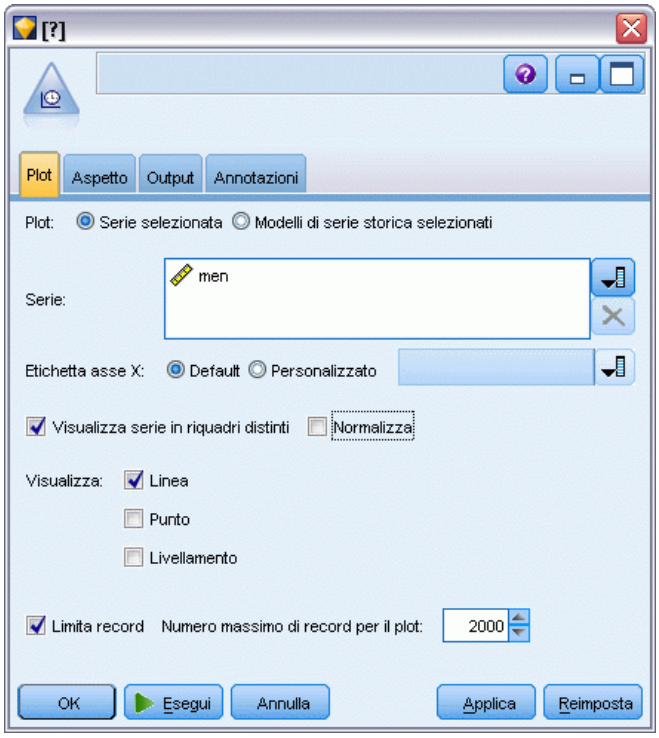
- Aprire il nodo File di input IBM® SPSS® Statistics e selezionare la scheda Tipi.
- Fare clic su Leggi valori, quindi su OK.
- Fare clic sulla colonna *Ruolo* relativa al campo *uomini* e impostare il ruolo su Obiettivo.
- Impostare il ruolo di tutti gli altri campi su Nessuno e fare clic su OK.

Figura 15-3
Impostazione dell'intervallo di tempo



- Collegare un nodo Intervalli di tempo al nodo File di input SPSS Statistics.
- Aprire il nodo Intervalli di tempo e impostare Intervallo di tempo su Mesi.
- Selezionare Crea dai dati.
- Impostare Campo su data e fare clic su OK.

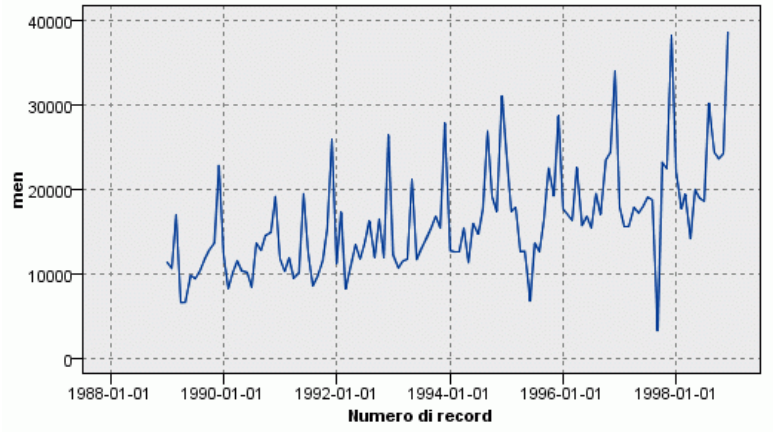
Figura 15-4
Rappresentazione grafica della serie storica



- Collegare un nodo Plot tempo al nodo Intervalli di tempo.
- Nella scheda Plot, aggiungere uomini all'elenco Serie.
- Deselezionare la casella di controllo Normalizza.
- Fare clic su Esegui.

Esame dei dati

Figura 15-5
Vendite effettive di abbigliamento maschile



La serie mostra un trend generale al rialzo; vale a dire, i valori della serie tendono ad aumentare con il passare del tempo. Il trend al rialzo appare costante, pertanto si tratta di un trend lineare.

La serie inoltre presenta un andamento stagionale chiaro, con picchi annuali nel mese di dicembre, come indicano le linee verticali del grafico. Le variazioni stagionali sembrano aumentare parallelamente al trend rialzista della serie, il che suggerisce una stagionalità moltiplicativa piuttosto che additiva.

▶ Fare clic su OK per chiudere la rappresentazione grafica.

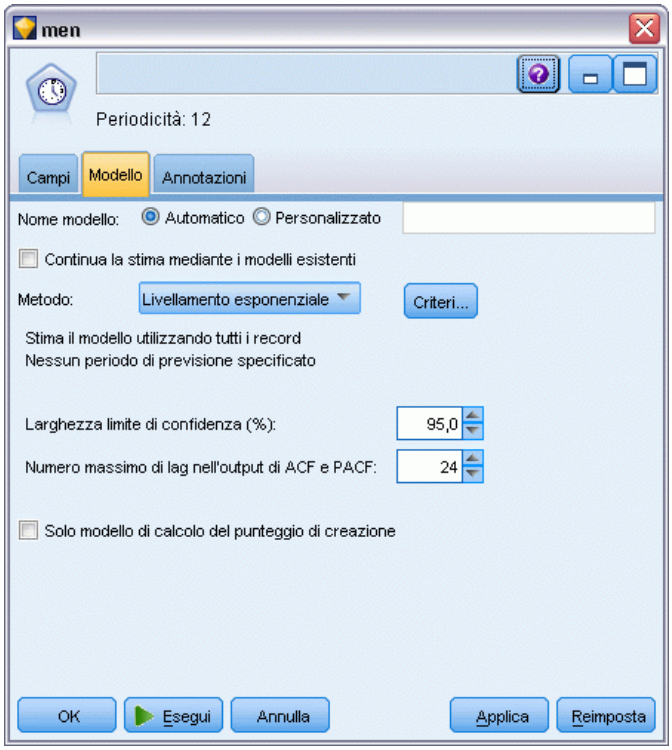
Dopo che sono state identificate le caratteristiche dalla serie, è possibile cominciare a eseguirne la modellazione. Il metodo di livellamento esponenziale è utile per calcolare previsioni di serie che evidenziano caratteristiche tendenziali, stagionali o entrambe. Come notato, i dati esaminati qui presentano entrambe le caratteristiche.

Livellamento esponenziale

La creazione del modello di livellamento esponenziale più adatto implica l'identificazione del tipo di modello, vale a dire se il modello deve includere trend, stagionalità o entrambe le caratteristiche, e la determinazione dei parametri più adatti al modello scelto.

La rappresentazione grafica delle vendite di abbigliamento maschile nel corso del tempo suggerisce l'impiego di un modello con una componente di trend lineare e una componente di stagionalità moltiplicativa. Queste caratteristiche implicano un modello di Winters. Tuttavia, si inizierà con l'esaminare un modello semplice (privo di trend e stagionalità), quindi un modello Holt (che incorpora un trend lineare ma nessuna stagionalità). In questo modo si apprenderà a identificare un modello non adatto ai dati, competenza fondamentale per una corretta creazione di modelli.

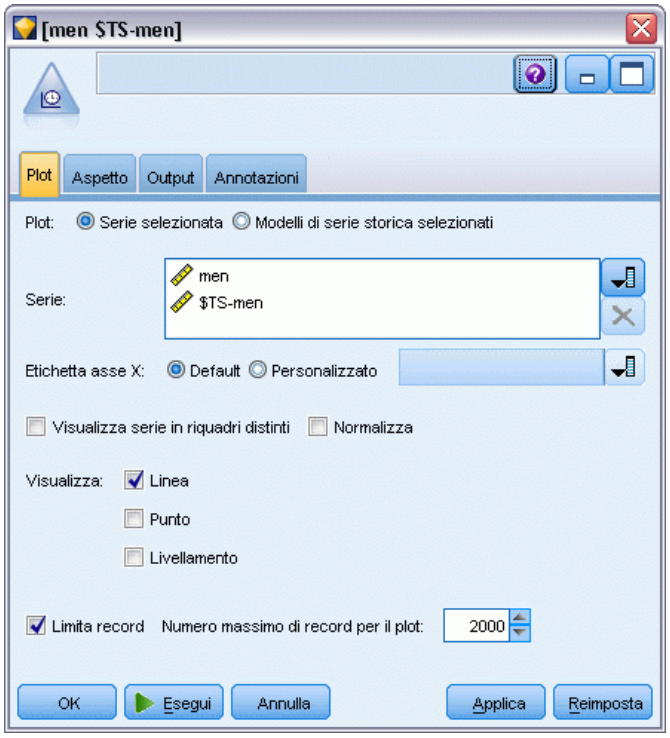
Figura 15-6
Specifica del livellamento esponenziale



Si inizierà con un modello di livellamento esponenziale semplice.

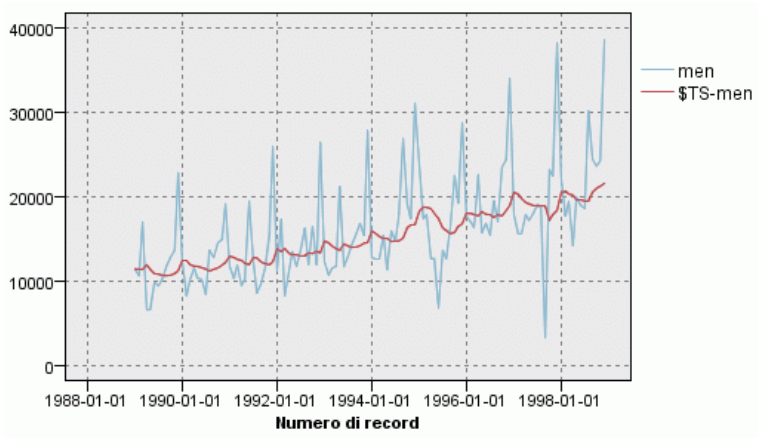
- Collegare un nodo Serie storica al nodo Intervalli di tempo.
- Nella scheda Modello, impostare Metodo su Livellamento esponenziale.
- Fare clic su Esegui per creare l'insieme di modelli.

Figura 15-7
Rappresentazione grafica del modello di serie storica



- Collegare un nodo Plot tempo all'insieme di modelli.
- Nella scheda Plot, aggiungere *uomini* e *$TS-uomini* all'elenco Serie.
- Deselezionare le caselle di controllo Visualizza serie in riquadri distinti e Normalizza.
- Fare clic su Esegui.

Figura 15-8
Modello di livellamento esponenziale semplice

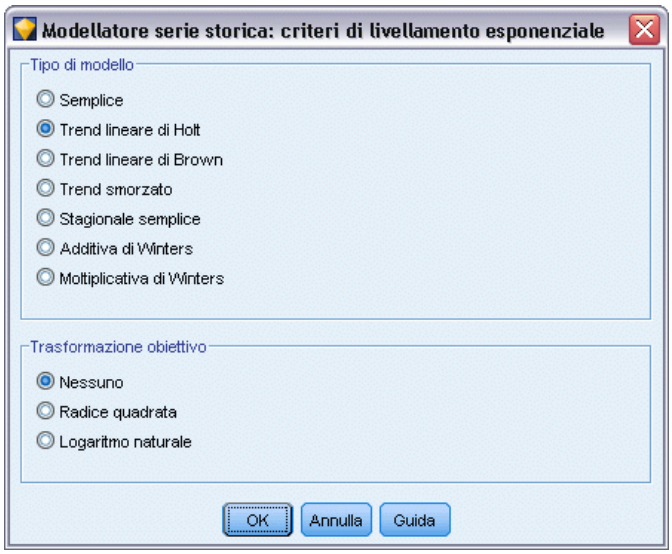


Il plot uomini rappresenta i dati effettivi mentre $TS-uomini indica il modello di serie storica.

Il modello semplice evidenzia un trend al rialzo graduale (e piuttosto massiccio), ma non tiene conto della stagionalità. Questo modello può essere quindi scartato.

- Fare clic su OK per chiudere la finestra del plot tempo.

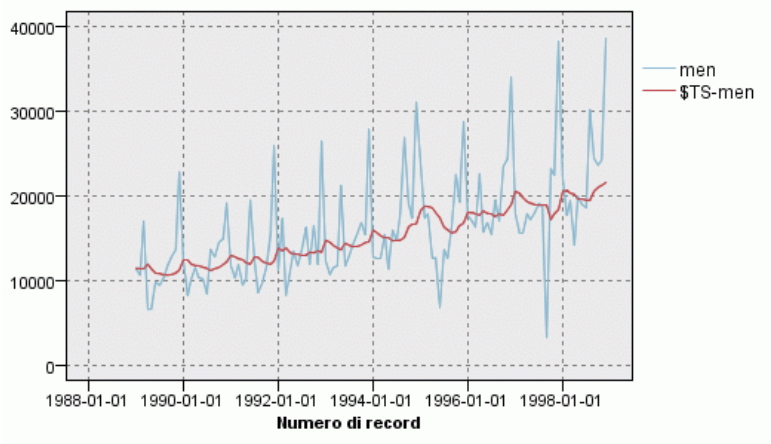
Figura 15-9
Selezione del modello di Holt



Si procederà ora alla creazione del modello lineare di Holt che, probabilmente, non sarà in grado di riprodurre la stagionalità, ma dovrebbe consentire di modellare il trend meglio del modello semplice.

- Riaprire il nodo Serie storica.
- Nella scheda Modello, con il metodo Livellamento esponenziale ancora selezionato, fare clic su Criteri.
- Nella finestra di dialogo Criteri di livellamento esponenziale, scegliere Trend lineare di Holt.
- Fare clic su OK per chiudere la finestra di dialogo.
- Fare clic su Esegui per ricreare l'insieme di modelli.
- Riaprire il nodo Plot tempo e fare clic su Esegui.

Figura 15-10
Modello a trend lineare di Holt

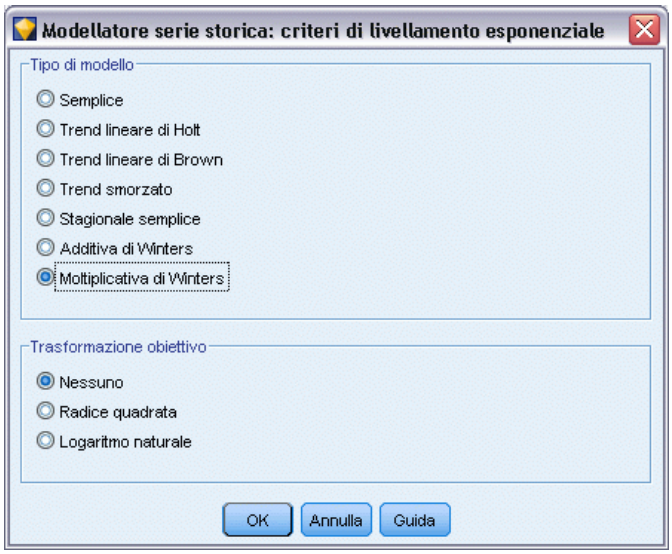


Il modello di Holt visualizza un trend al rialzo più lineare che nel caso del modello semplice, ma ancora non tiene conto della stagionalità; pertanto, può essere scartato.

▶ Chiudere la finestra del plot tempo.

Si ricorderà che la rappresentazione grafica iniziale delle vendite di abbigliamento maschile nel corso del tempo suggeriva l'utilizzo di un modello con una componente di trend lineare e una componente di stagionalità moltiplicativa. Pertanto, il modello di Winters potrebbe essere più indicato.

Figura 15-11
Selezione del modello di Winters

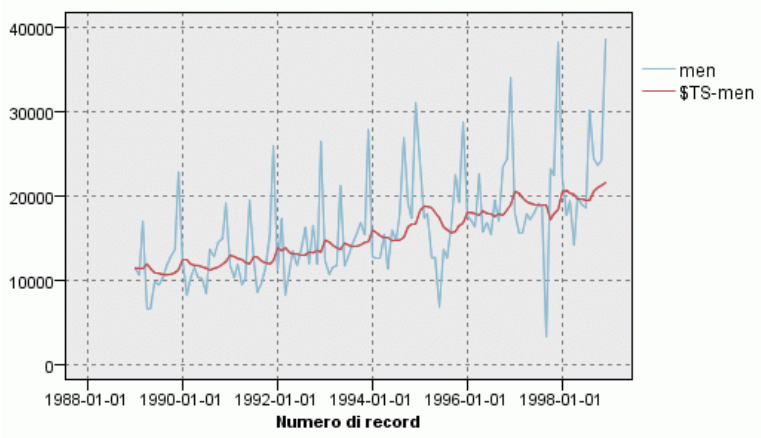


▶ Riaprire il nodo Serie storica.

▶ Nella scheda Modello, con il metodo Livellamento esponenziale ancora selezionato, fare clic su Criteri.

- Nella finestra di dialogo Criteri di livellamento esponenziale, scegliere Moltiplicativa di Winters.
- Fare clic su OK per chiudere la finestra di dialogo.
- Fare clic su Esegui per ricreare l'insieme di modelli.
- Aprire il nodo Plot tempo e fare clic su Esegui.

Figura 15-12
Modello moltiplicativo di Winters



Il modello appare più adeguato in quanto riflette sia il trend che la stagionalità dei dati.

L'insieme di dati copre un periodo di 10 anni e mostra 10 picchi stagionali in corrispondenza del mese di dicembre di ogni anno. I 10 picchi presenti nei risultati previsti corrispondono ai 10 picchi annuali nei dati reali.

Tuttavia, i risultati evidenziano anche i limiti della procedura di livellamento esponenziale. Osservando i picchi verso l'alto e verso il basso, appare evidente una struttura significativa che non viene spiegata.

Se si è principalmente interessati alla modellazione di una tendenza di lungo termine con variazioni stagionali, il metodo del livellamento esponenziale può essere adeguato. Tuttavia, se si desidera modellare una struttura più complessa, come quella presa in considerazione qui, è consigliabile provare a utilizzare la procedura ARIMA.

ARIMA

Questa procedura consente di creare un modello ARIMA (modello autoregressivo integrato a media mobile), adatto alla creazione di modelli di serie storiche più accurati. I modelli ARIMA offrono metodi più sofisticati per la modellazione delle componenti di trend e stagionalità rispetto ai modelli di livellamento esponenziale, con l'ulteriore vantaggio di includere le variabili predittore.

Proseguendo con l'esempio della società di vendite mediante catalogo che desidera sviluppare un modello di previsione, si è osservato che la società ha raccolto i dati di vendita mensili di abbigliamento maschile insieme a diverse serie che possono essere utilizzate per spiegare alcune delle variazioni nell'andamento delle vendite. I predittori possibili includono il numero di cataloghi inviati per posta e il numero delle pagine nel catalogo, il numero di linee telefoniche

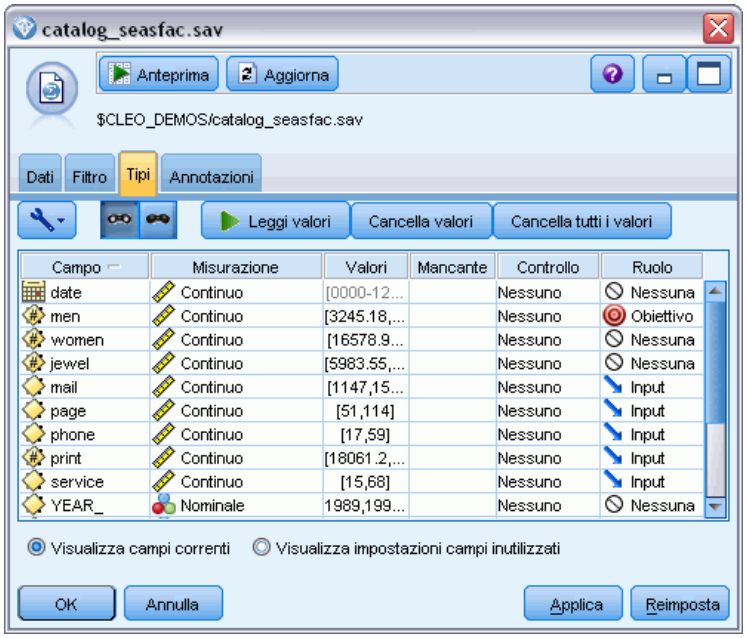
aperte per le ordinazioni, la spesa sostenuta per la pubblicità a mezzo stampa e il numero dei rappresentanti del servizio di assistenza clienti.

Uno o più di tali predittori sono utili per le previsioni? Un modello che include i predittori è realmente migliore di un modello privo di predittori? Utilizzando la procedura ARIMA, è possibile creare un modello di previsione con predittori e valutare se esiste una differenza significativa nella capacità di previsione rispetto al modello di livellamento esponenziale privo di predittori.

Il metodo ARIMA consente di perfezionare il modello specificando gli ordini di autoregressione, differenziazione e media mobile, oltre agli equivalenti stagionali di questi componenti. La determinazione manuale dei valori migliori per questi componenti può essere un'operazione piuttosto lunga e complessa, pertanto, ai fini di questo esempio, sarà Expert Modeler a scegliere un modello ARIMA.

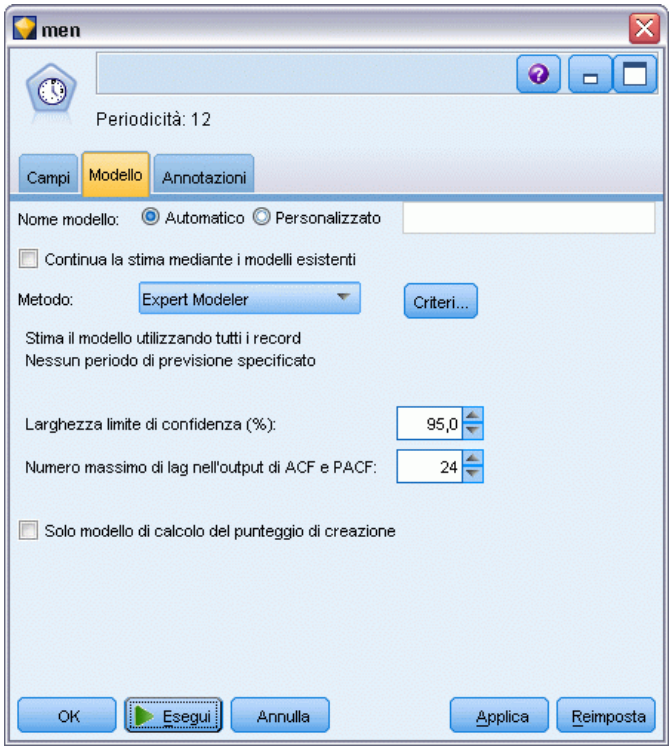
Si cercherà quindi di creare un modello migliore considerando alcune delle altre variabili nell'insieme di dati come variabili predittore. I predittori che sembrano più utili sono il numero di cataloghi inviati per posta (*posta*), il numero delle pagine nel catalogo (*pagina*), il numero di linee telefoniche aperte per le ordinazioni (*telefono*), la spesa sostenuta per le pubblicità a mezzo stampa (*stampa*) e il numero dei rappresentanti del servizio di assistenza clienti (*servizio*).

Figura 15-13
Impostazione dei campi predittori



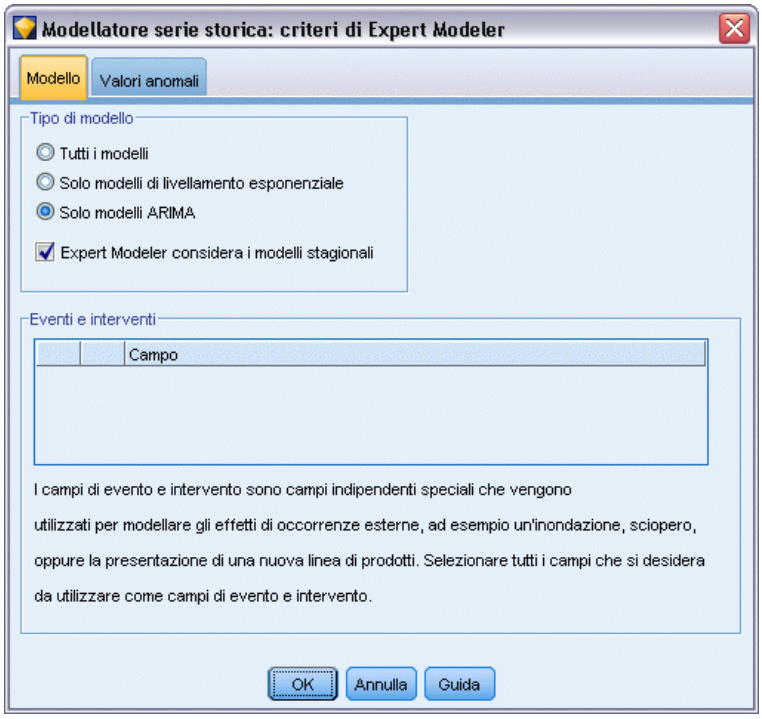
- Aprire il nodo File di input IBM® SPSS® Statistics.
- Nella scheda Tipi, impostare il *Ruolo* di *posta*, *pagina*, *telefono*, *stampa* e *servizio* su Input.
- Verificare che il ruolo di uomini sia impostato su Obiettivo e che tutti i campi rimanenti siano impostati su Nessuno.
- Fare clic su OK.

Figura 15-14
Scelta di Expert Modeler



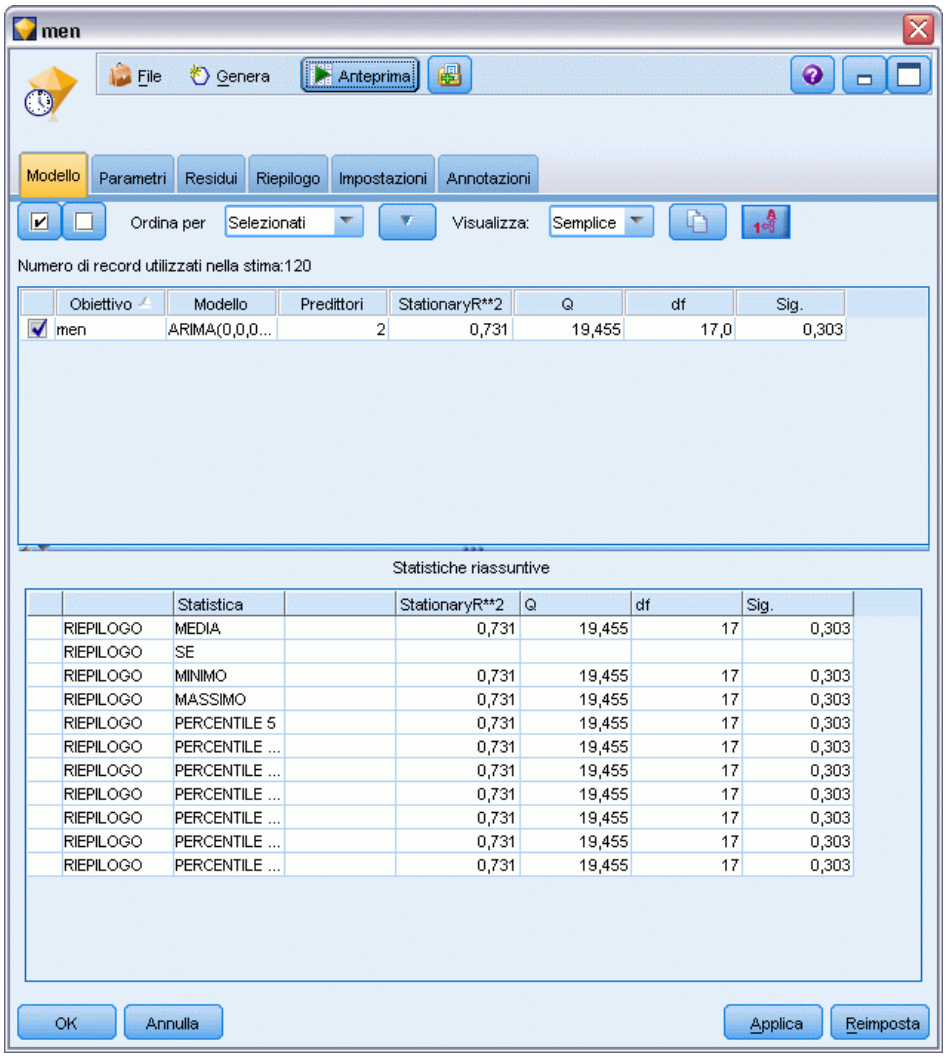
- Aprire il nodo Serie storica.
- Nella scheda Modello, impostare Metodo su Expert Modeler e fare clic su Criteri.

Figura 15-15
Scelta dei soli modelli ARIMA



- Nella finestra di dialogo Criteri di Expert Modeler, scegliere l'opzione Solo modelli ARIMA e verificare che sia selezionata l'opzione Expert Modeler considera i modelli stagionali.
- Fare clic su OK per chiudere la finestra di dialogo.
- Fare clic su Esegui nella scheda Modello per ricreare l'insieme di modelli.

Figura 15-16
Expert Modeler sceglie due predittori



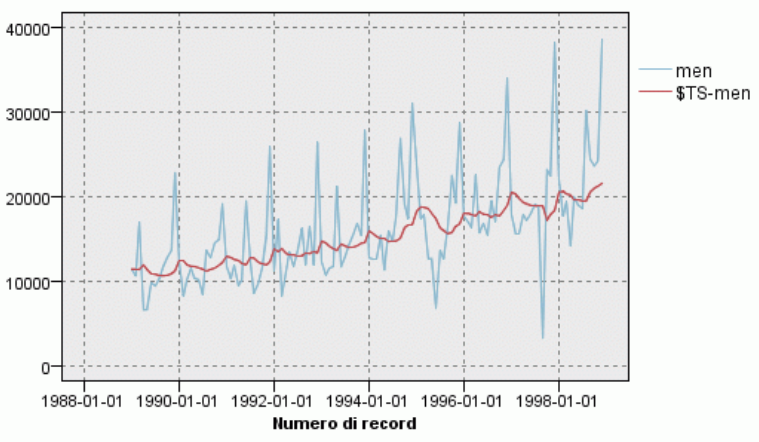
- Aprire l'insieme di modelli.

  Si noti che Expert Modeler ha scelto come significativi soltanto due dei cinque predittori specificati.

- Fare clic su OK per chiudere l'insieme di modelli.

- Aprire il nodo Plot tempo e fare clic su Esegui.

Figura 15-17
Modello ARIMA con predittori specificati



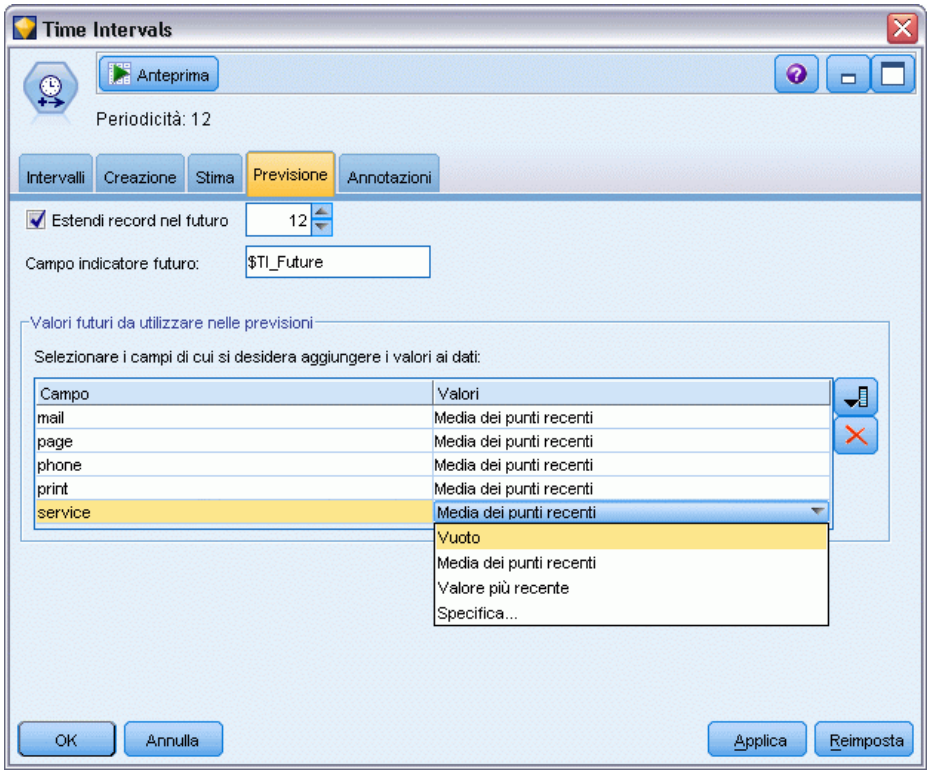
Questo modello è migliore rispetto a quello precedente in quanto coglie anche il grande picco discendente; per questo risulta il modello più adatto fra quelli presi in considerazione finora.

A questo punto, è possibile cercare di perfezionare ulteriormente il modello, ma qualsiasi miglioramento sarebbe di entità minima. Si è stabilito quindi che il modello ARIMA contenente i predittori è il modello preferibile, pertanto verrà utilizzato il modello appena creato. Ai fini di questo esempio, verrà effettuata una previsione delle vendite per l'anno prossimo.

- Fare clic su OK per chiudere la finestra del plot tempo.
- Aprire il nodo Intervalli di tempo e selezionare la scheda *Previsione*.
- Selezionare la casella di controllo *Estendi record nel futuro* e impostare il relativo valore su 12.

L'uso dei predittori nel calcolo delle previsioni richiede la specifica dei valori stimati per quei campi nel periodo di previsione, in modo che il modellatore possa prevedere il campo obiettivo con maggior precisione.

Figura 15-18
Specifica dei valori futuri per i campi predittori

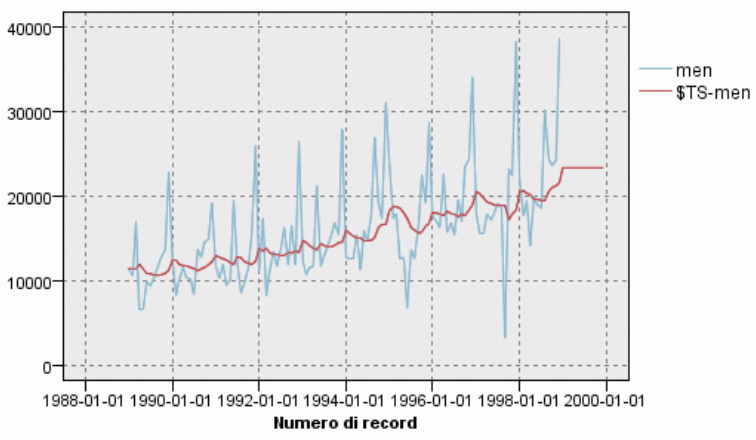


- Nel gruppo Valori futuri da utilizzare nelle previsioni, fare clic sul pulsante di selezione dei campi posto a destra della colonna Valori.
- Nella finestra di dialogo Seleziona campi, selezionare i campi da posta a servizio e fare clic su OK.

  In una situazione reale, a questo punto verrebbero specificati i valori futuri manualmente, in quanto i cinque predittori si riferiscono a fattori sotto il controllo dell'utente. Ai fini di questo esempio, verrà invece utilizzata una delle funzioni predefinite per non dover specificare 12 valori per ogni predittore. Dopo avere acquisito maggior familiarità con questo esempio, provare a utilizzare diversi valori futuri per verificare che effetto hanno sul modello.
- Per ogni campo, fare clic sul campo Valori in modo da visualizzare un elenco dei valori possibili, quindi scegliere Media dei punti recenti. Questa opzione calcola la media degli ultimi tre punti di dati per questo campo e la utilizza come valore stimato per ciascun caso.
- Fare clic su OK.
- Aprire il nodo Serie storica e fare clic su Esegui per ricreare l'insieme di modelli.
- Aprire il nodo Plot tempo e fare clic su Esegui.

Le previsioni per il 1999 appaiono buone; in linea con le aspettative, si registra un ritorno ai normali livelli di vendita dopo il picco di dicembre e un trend stabilmente al rialzo nella seconda metà dell'anno, con vendite decisamente superiori a quelle dell'anno precedente.

Figura 15-19
Previsioni di vendita con predittori specificati



Riepilogo

È stato correttamente creato il modello di una serie storica complessa, che incorpora non solo un trend al rialzo ma anche variazioni stagionali e di altro tipo. Inoltre, si è visto come, mediante una serie di tentativi, è possibile creare un modello sempre più accurato, che in seguito è stato utilizzato per prevedere i volumi di vendita futuri.

In pratica, sarà necessario applicare nuovamente il modello man mano che i dati sulle vendite effettive vengono aggiornati (per esempio ogni mese oppure ogni trimestre) e produrre previsioni aggiornate. Per ulteriori informazioni, vedere l'argomento Riapplicazione di un modello di serie storica in il capitolo 14 a pag. 197.

Capitolo
16

Offerte ai clienti (Autoapprendimento)

Il nodo Modello risposta autoapprendimento (SLRM) genera e attiva l'aggiornamento di un modello che consente di individuare le offerte più indicate per i clienti e prevedere la probabilità che vengano accettate. Questi tipi di modelli sono i più utili nell'ambito del CRM (Customer Relationship Management), per esempio nelle applicazioni di call center o marketing.

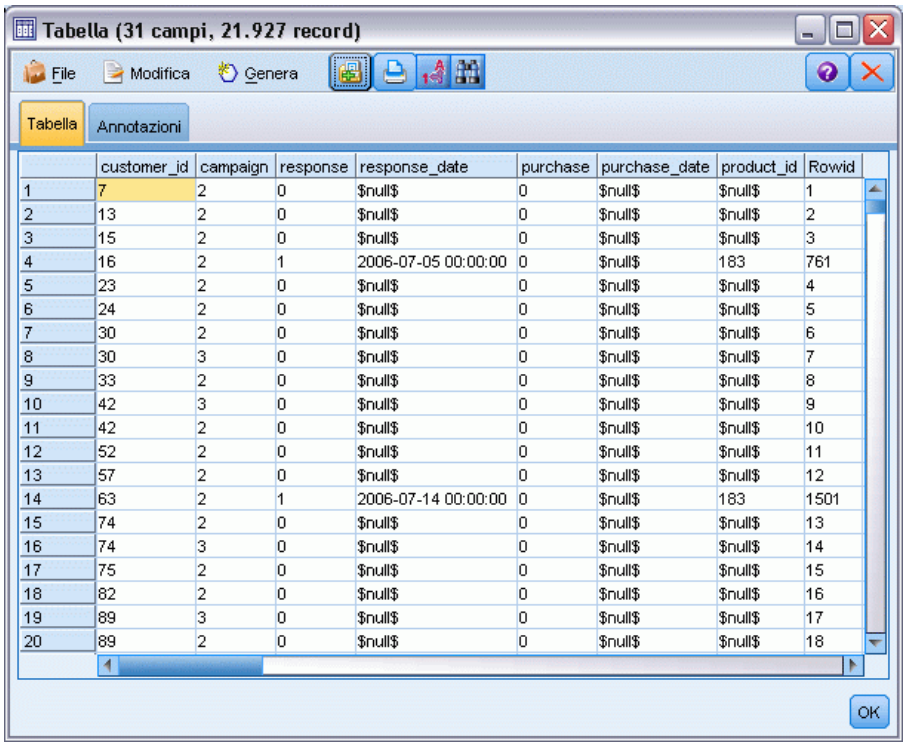
Questo esempio si basa su un gruppo bancario fittizio. La divisione di marketing desidera ottenere risultati più redditizi nelle campagne future, inviando offerte mirate di servizi finanziari ai singoli clienti. Specificatamente, l'esempio utilizza un Modello risposta autoapprendimento per identificare le caratteristiche dei clienti che, stando ai risultati delle offerte precedenti, hanno più probabilità di rispondere in modo favorevole e per promuovere le migliori offerte attualmente disponibili in base ai risultati ottenuti.

Questo esempio utilizza lo stream *pm_selflearn.str*, che fa riferimento ai file di dati *pm_customer_train1.sav*, *pm_customer_train2.sav* e *pm_customer_train3.sav*. Questi file sono disponibili nella cartella *Demos* di tutte le installazioni di IBM® SPSS® Modeler. L'accesso si effettua dal gruppo di programmi IBM® SPSS® Modeler nel menu Start di Windows. Il file *pm_selflearn.str* si trova nella cartella *streams*.

Dati esistenti

La società dispone dei dati storici che tengono traccia delle offerte fatte ai clienti nell'ambito delle campagne precedenti e delle risposte a tali offerte. Questi dati comprendono anche le informazioni finanziarie e demografiche che possono essere utilizzate per prevedere i tassi di risposta dei diversi clienti.

Figura 16-1
Risposte alle offerte precedenti

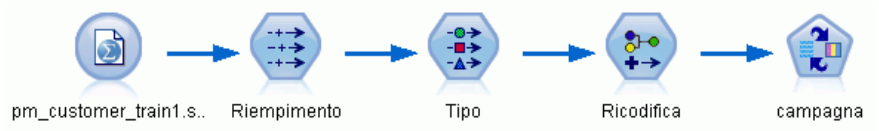


	customer_id	campaign	response	response_date	purchase	purchase_date	product_id	Rowid
1	7	2	0	$null$	0	$null$	$null$	1
2	13	2	0	$null$	0	$null$	$null$	2
3	15	2	0	$null$	0	$null$	$null$	3
4	16	2	1	2006-07-05 00:00:00	0	$null$	183	761
5	23	2	0	$null$	0	$null$	$null$	4
6	24	2	0	$null$	0	$null$	$null$	5
7	30	2	0	$null$	0	$null$	$null$	6
8	30	3	0	$null$	0	$null$	$null$	7
9	33	2	0	$null$	0	$null$	$null$	8
10	42	3	0	$null$	0	$null$	$null$	9
11	42	2	0	$null$	0	$null$	$null$	10
12	52	2	0	$null$	0	$null$	$null$	11
13	57	2	0	$null$	0	$null$	$null$	12
14	63	2	1	2006-07-14 00:00:00	0	$null$	183	1501
15	74	2	0	$null$	0	$null$	$null$	13
16	74	3	0	$null$	0	$null$	$null$	14
17	75	2	0	$null$	0	$null$	$null$	15
18	82	2	0	$null$	0	$null$	$null$	16
19	89	3	0	$null$	0	$null$	$null$	17
20	89	2	0	$null$	0	$null$	$null$	18

Creazione dello stream

- Aggiungere un nodo di input File Statistics che punti al file *pm_customer_train1.sav* situato nella directory *Demos* dell'installazione IBM® SPSS® Modeler.

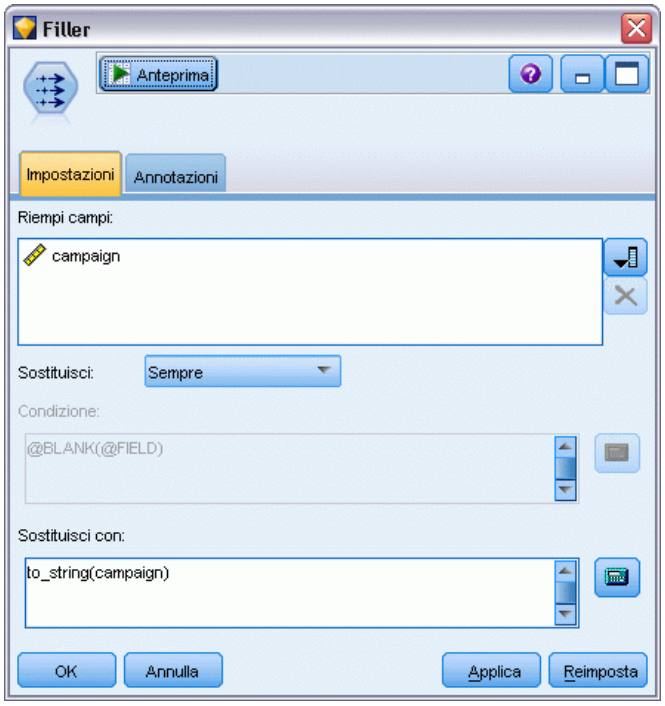
Figura 16-2
Stream campione SLRM



- Aggiungere un nodo Riempimento e selezionare campagna come campo di riempimento.
- Selezionare un tipo Sostituisci di Sempre.

▶ Nella casella di testo Sostituisci con, immettere to_string(campaign) e fare clic su OK.

Figura 16-3
Nuovo campo campagna



▶ Aggiungere un nodo Tipo e impostare il *Ruolo* su Nessuno per i campi *id_cliente*, *data_risposta*, *data_acquisto*, *id_prodotto*, *IDriga* e *X_casuale*.

Figura 16-4
Modifica delle impostazioni del nodo Tipo

- Impostare il *Ruolo* su Obiettivo per i campi *campagna* e *risposta*. Questi sono i campi su cui si baseranno le previsioni.

  Impostare la Misurazione su Flag per il campo *risposta*.

- Fare clic su Leggi valori, quindi su OK.

  Poiché i dati del campo campagna mostrano un elenco di numeri (1, 2, 3 e 4), è possibile riclassificare i campi per disporre di titoli più significativi.

- Aggiungere un nodo Ricodifica al nodo Tipo.

- Nel campo Ricodifica, selezionare Campo esistente.

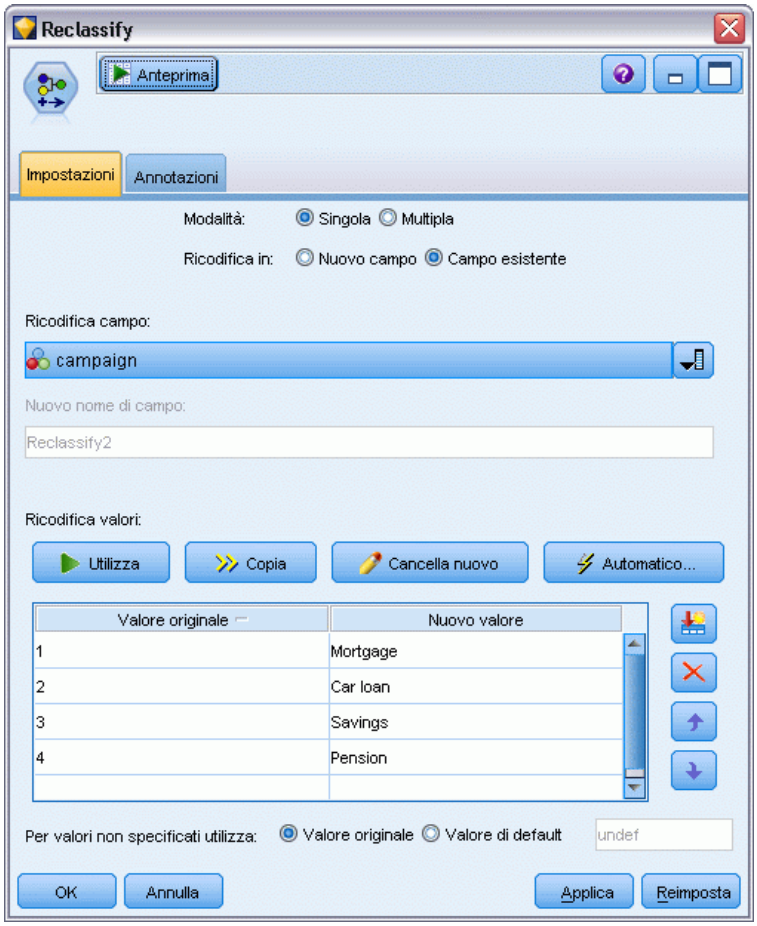
- Nell'elenco Campo Ricodifica, selezionare campagna.

- Fare clic sul pulsante Recupera; i valori del campo campagna vengono aggiunti alla colonna *Valore originale*.

- Nella colonna *Nuovo valore*, immettere i seguenti nomi di campagna nelle prime quattro righe:
  - Prestito ipotecario
  - Mutuo auto
  - Risparmio
  - Pensione

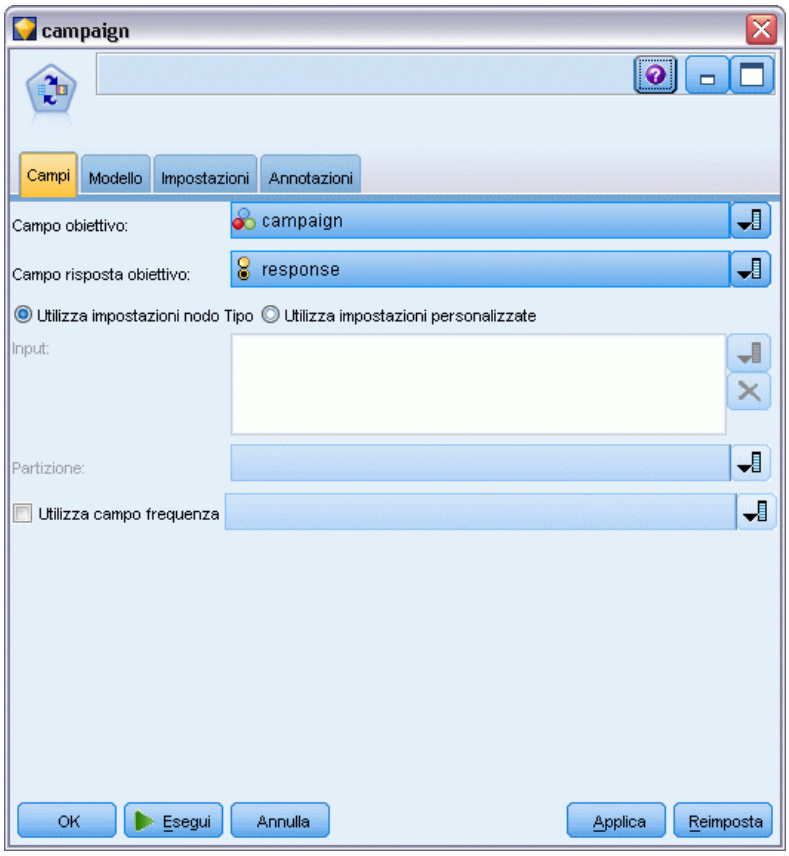
▶ Fare clic su OK.

Figura 16-5
Ricodificazione dei nomi di campagna



- Associare un nodo Modelli SLRM al nodo Ricodifica. Nella scheda Campi, selezionare campagna per il campo Obiettivo e risposta per il campo Risposta obiettivo.

Figura 16-6
Selezione di Obiettivo e Risposta obiettivo

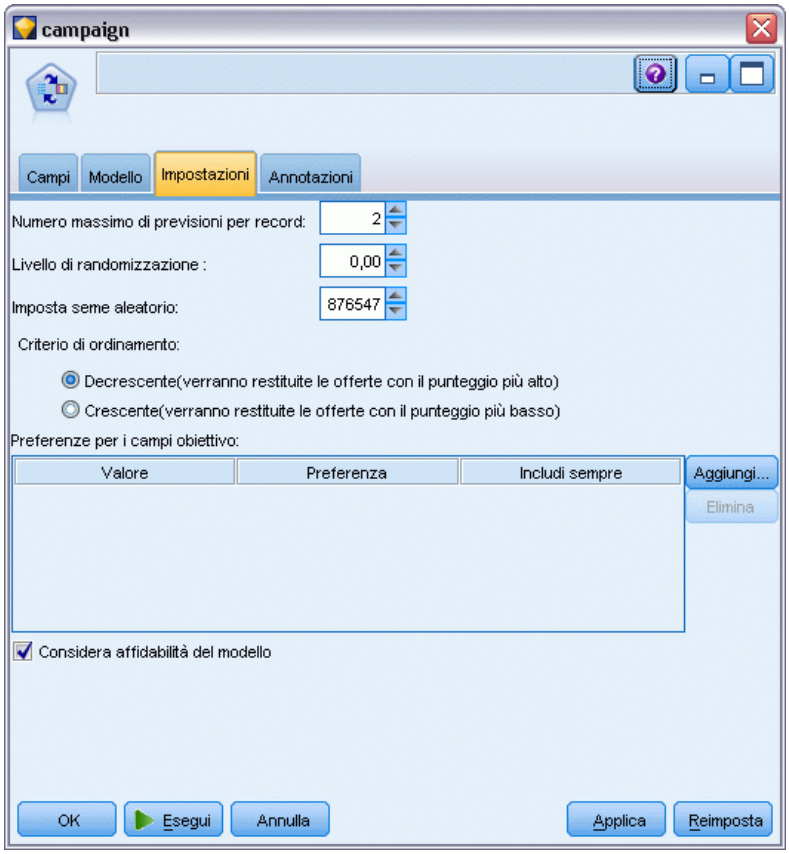


- Nel campo Numero massimo di previsioni per record della scheda Impostazioni, ridurre a 2 il valore.

  Ciò significa che per ogni cliente vi saranno due offerte identificate che avranno la massima probabilità di venire accettate.

▶ Assicurarsi che sia selezionata l'opzione Considera affidabilità del modello e fare clic su Esegui.

Figura 16-7
Impostazioni del nodo SLRM



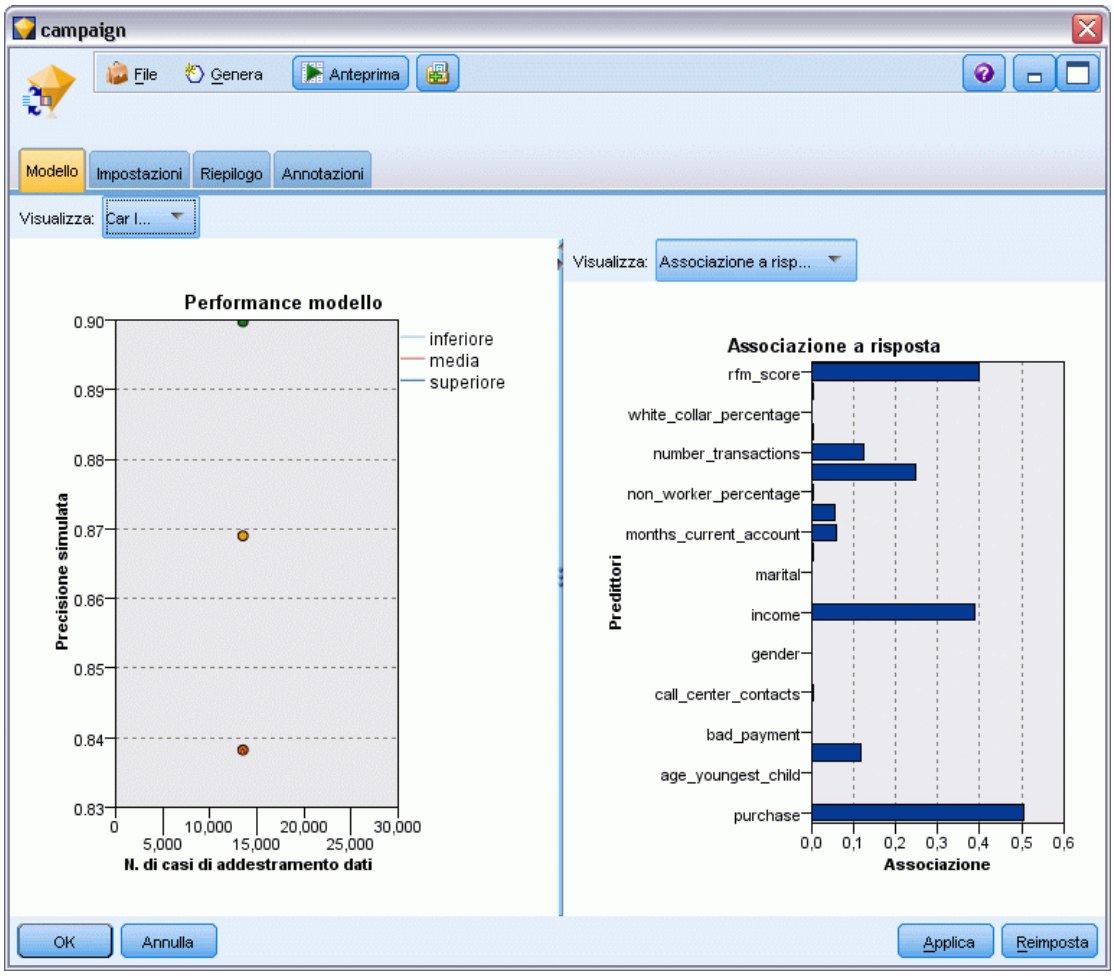
Visualizzazione del modello

▶ Aprire l'insieme di modelli. All'inizio, la scheda Modello mostra la precisione stimata delle previsioni per ogni offerta e l'importanza relativa di ogni predittore nella stima del modello.

Per visualizzare la correlazione dei singoli predittori con la variabile obiettivo, scegliere Associazione a risposta dall'elenco Visualizza nel riquadro di destra.

▶ Per alternare tra le quattro offerte per le quali sono disponibili previsioni, selezionare l'offerta desiderata nell'elenco Visualizza nel riquadro di sinistra.

Figura 16-8
Insieme di modelli SLRM

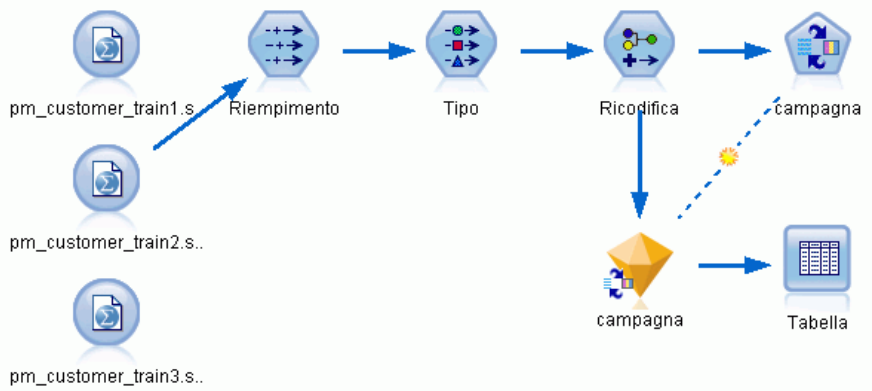


▶ Chiudere la finestra dell'insieme di modelli.

▶ Nell'area di disegno dello stream, disconnettere il nodo di input File IBM® SPSS® Statistics che punta a *pm_customer_train1.sav*.

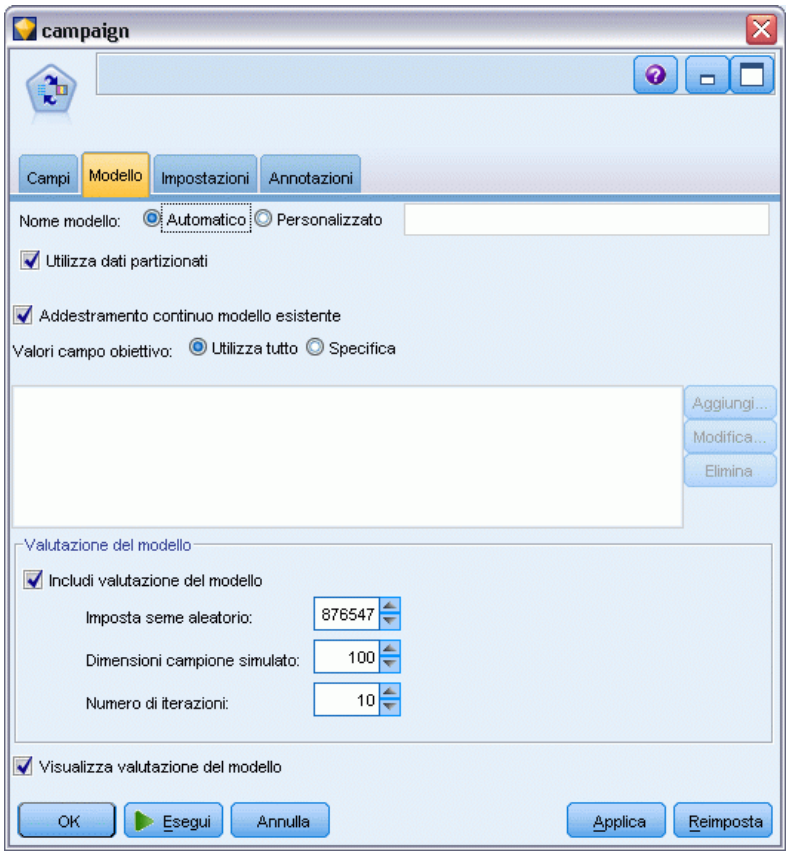
- Aggiungere un nodo di input File Statistics che punti al file *pm_customer_train2.sav* situato nella cartella *Demos* dell'installazione IBM® SPSS® Modeler, quindi connetterlo al nodo Riempimento.

Figura 16-9
Associazione di una seconda sorgente dati allo stream SLRM



- Nella scheda Modello del nodo SLRM, selezionare Addestramento continuo modello esistente.

Figura 16-10
Addestramento continuo del modello

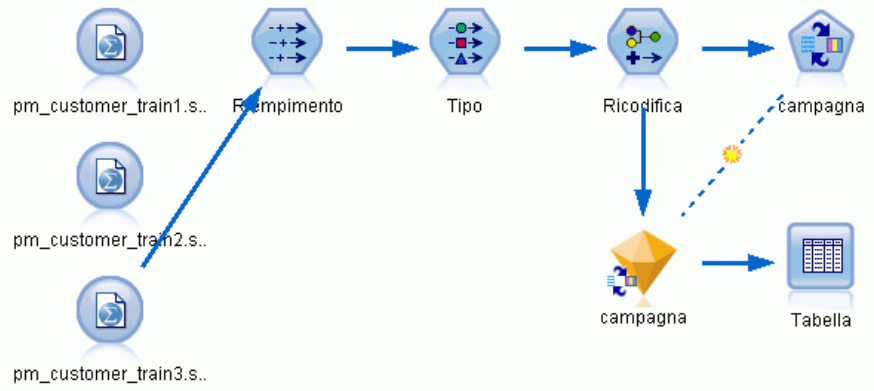


- Fare clic su Esegui per ricreare l'insieme di modelli. Per visualizzare i relativi dettagli, fare doppio clic sull'insieme di modelli nell'area di disegno.

  Nella scheda Modello sono ora visualizzate le stime riviste della precisione delle previsioni per ogni offerta.

- Aggiungere un nodo di input File Statistics che punti al file *pm_customer_train3.sav* situato nella cartella *Demos* dell'installazione SPSS Modeler, quindi connetterlo al nodo Riempimento.

Figura 16-11
Associazione di una terza sorgente dati allo stream SLRM

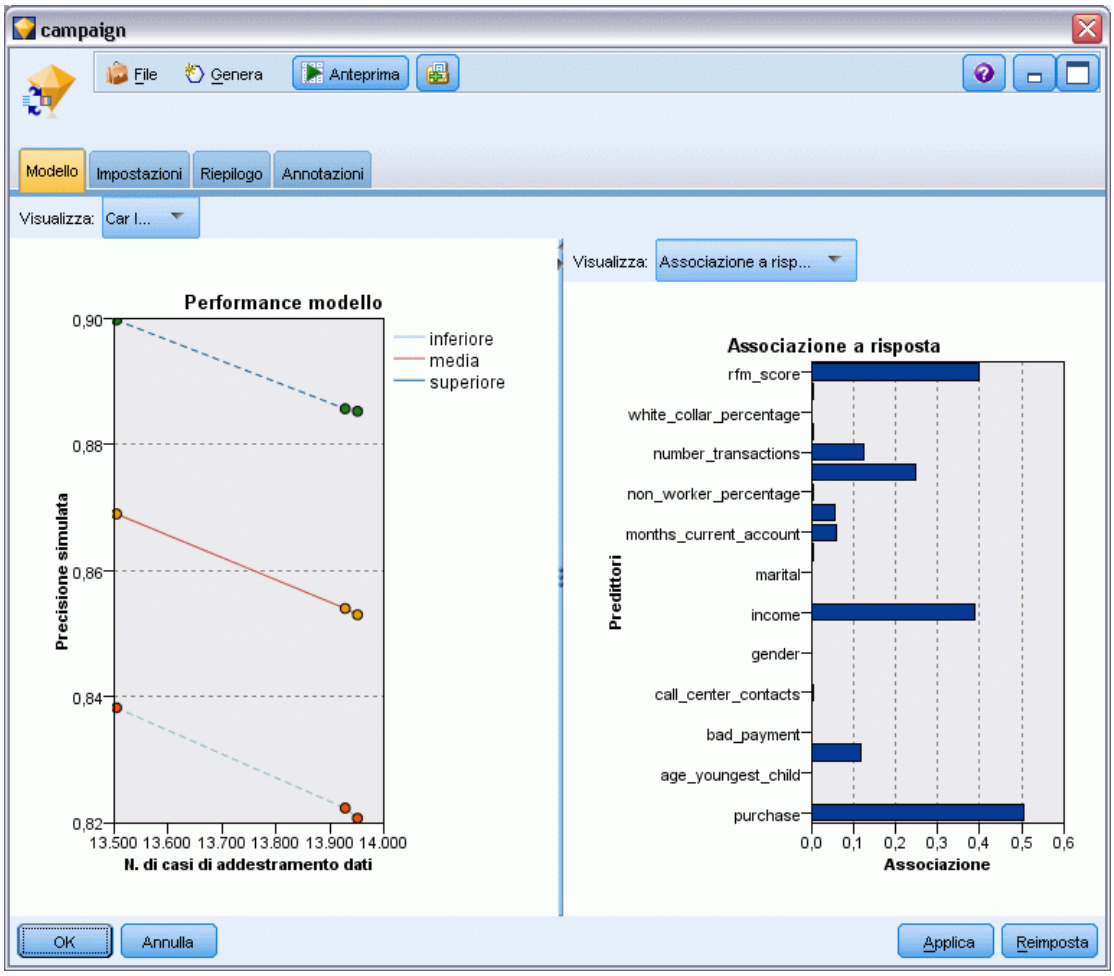


- Fare clic su Esegui per ricreare ancora una volta l'insieme di modelli. Per visualizzare i relativi dettagli, fare doppio clic sull'insieme di modelli nell'area di disegno.

- Nella scheda Modello è ora visualizzata la precisione stimata finale delle previsioni per ogni offerta.

Come si può notare, la precisione media è diminuita leggermente con l'aggiunta di ulteriori sorgenti dati (da 86,9% a 85,4%); tuttavia, tale fluttuazione è minima e potrebbe essere imputata a lievi anomalie nei dati disponibili.

Figura 16-12
Insieme di modelli SLRM aggiornato



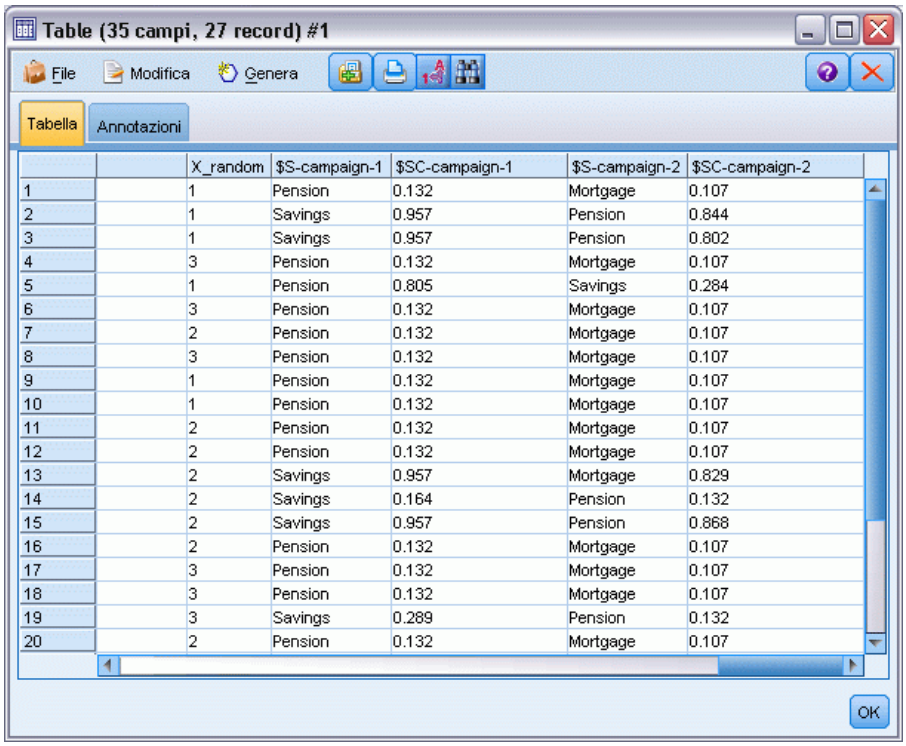
- Associare un nodo Tabella all'ultimo (terzo) modello generato ed eseguirlo.

- Scorrere per visualizzare la parte destra della tabella. Le previsioni indicano quali offerte è più probabile che vengano accettate da un cliente e la confidenza che tali clienti accetteranno, in base ai dettagli di ogni cliente.

  Per esempio, la prima riga della tabella mostra il tasso di confidenza secondo cui un cliente che ha acceso un mutuo sull'auto accetterà un piano pensione se gliene verrà offerto uno è pari solo al 13,2% (come mostra il valore 0,132 nella colonna *$SC-campaign-1*). Tuttavia, le seconda e la terza riga mostrano altri due clienti che hanno acceso un mutuo sull'auto e in questi casi il tasso di confidenza che tali clienti, e altri clienti con storie analoghe, aprano un conto di risparmio è

pari al 95,7%, se gliene verrà offerto uno, e il tasso di confidenza che accettino una pensione è pari a oltre l'80%.

Figura 16-13
Output del modello: offerte e confidenze previste



		X_random	$S-campaign-1	$SC-campaign-1	$S-campaign-2	$SC-campaign-2
1		1	Pension	0.132	Mortgage	0.107
2		1	Savings	0.957	Pension	0.844
3		1	Savings	0.957	Pension	0.802
4		3	Pension	0.132	Mortgage	0.107
5		1	Pension	0.805	Savings	0.284
6		3	Pension	0.132	Mortgage	0.107
7		2	Pension	0.132	Mortgage	0.107
8		3	Pension	0.132	Mortgage	0.107
9		1	Pension	0.132	Mortgage	0.107
10		1	Pension	0.132	Mortgage	0.107
11		2	Pension	0.132	Mortgage	0.107
12		2	Pension	0.132	Mortgage	0.107
13		2	Savings	0.957	Mortgage	0.829
14		2	Savings	0.164	Pension	0.132
15		2	Savings	0.957	Pension	0.868
16		2	Pension	0.132	Mortgage	0.107
17		3	Pension	0.132	Mortgage	0.107
18		3	Pension	0.132	Mortgage	0.107
19		3	Savings	0.289	Pension	0.132
20		2	Pension	0.132	Mortgage	0.107

Per una spiegazione dei fondamenti matematici dei metodi di modellazione utilizzati in SPSS Modeler, vedere il manuale *SPSS Modeler Algorithms Guide*, disponibile nella directory *\Documentation* del DVD del prodotto.

Si noti inoltre che questi risultati si basano soltanto sui dati di addestramento. Per un'indicazione dell'efficacia con cui il modello potrà essere esteso ad altri dati in una situazione reale, utilizzare un nodo Partizione per estrarre un sottoinsieme di record da utilizzare per test e validazione. Per ulteriori informazioni, vedere l'argomento Nodo Partizione in il capitolo 4 in *IBM SPSS Modeler 15 Nodi di input, elaborazione e output*. Per ulteriori informazioni sul nodo SLRM, vedere il capitolo 14 nei Riferimenti ai nodi.

Capitolo
17

Previsione delle mancate restituzioni di prestiti (Rete bayesiana)

Le reti bayesiane consentono 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.

In questo esempio viene utilizzato lo stream denominato *bayes_bankloan.str*, che fa riferimento al file di dati *bankloan.sav*. Tali file sono disponibili nella directory *Demos* dell'installazione di IBM® SPSS® Modeler ed è possibile accedervi dal gruppo di programmi IBM® SPSS® Modeler nel menu Start di Windows. Il file *bayes_bankloan.str* si trova nella directory *streams*.

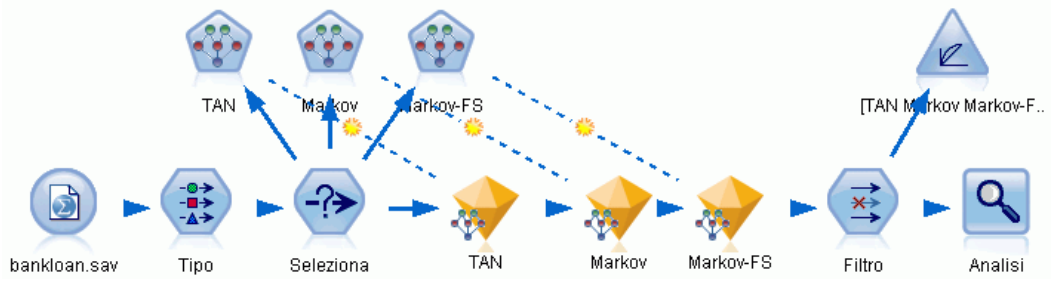
Si supponga, per esempio, che una banca desideri valutare le potenziali mancate restituzioni di prestiti. Potendo utilizzare dati precedenti sulle mancate restituzioni di prestiti per prevedere quali potenziali clienti potrebbero avere problemi di insolvenza, la banca potrebbe negare dei prestiti a questi clienti a rischio di credito negativo oppure offrire loro prodotti alternativi.

Questo esempio utilizza dati esistenti sulle mancate restituzioni di prestiti per prevedere i potenziali clienti insolventi futuri e considera tre diversi tipi di modello di rete bayesiana per stabilire qual è il migliore fra questi nel prevedere i potenziali clienti insolventi.

Creazione dello stream

- Aggiungere un nodo di input File Statistics che punta a *bankloan.sav* nella cartella *Demos*.

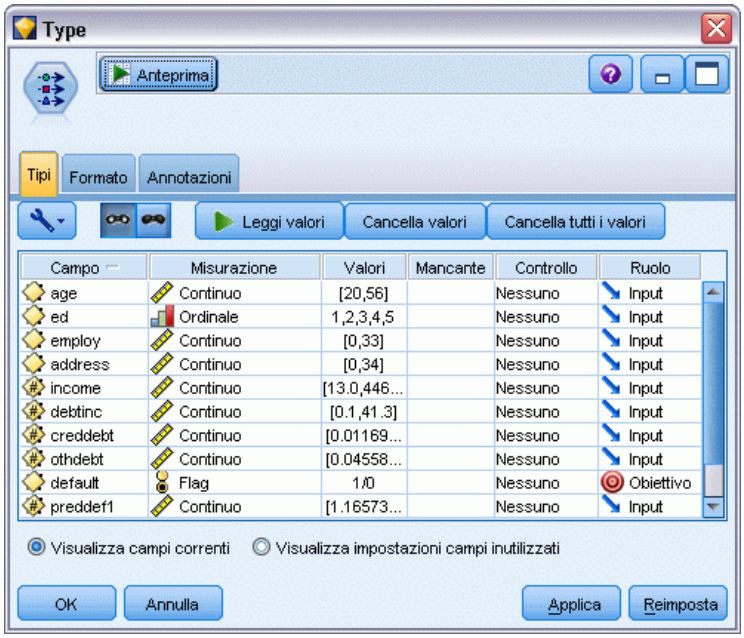
Figura 17-1
Stream di esempio di rete bayesiana



- Aggiungere un nodo Tipo al nodo di input e impostare il ruolo del campo di default su Obiettivo. Tutti gli altri campi dovrebbero avere il ruolo impostato su Input.

▶ Fare clic sul pulsante Leggi valori per popolare la colonna *Valori*.

Figura 17-2
Selezione del campo obiettivo



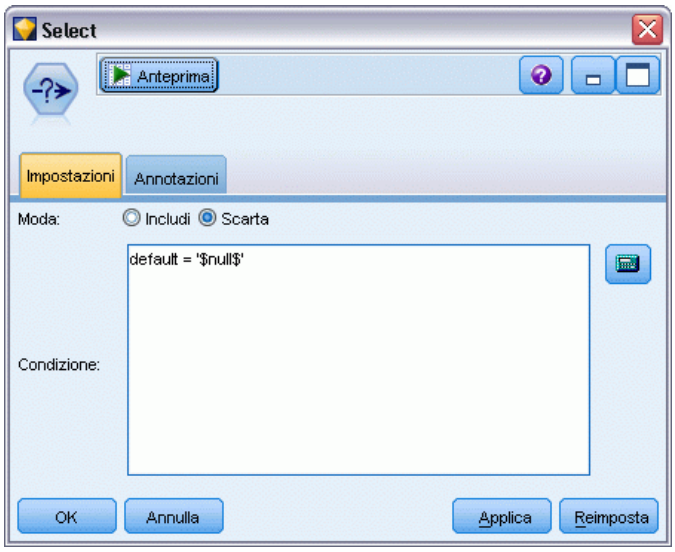
I casi in cui l'obiettivo ha un valore nullo non sono utilizzabili ai fini della generazione del modello. È possibile escluderli per evitare che vengano inclusi nella valutazione del modello.

▶ Aggiungere un nodo Seleziona al nodo Tipo.

▶ Per Modalità, selezionare Scarta.

▶ Nella casella Condizione, immettere default = '$null$'.

Figura 17-3
Eliminazione di obiettivi nulli



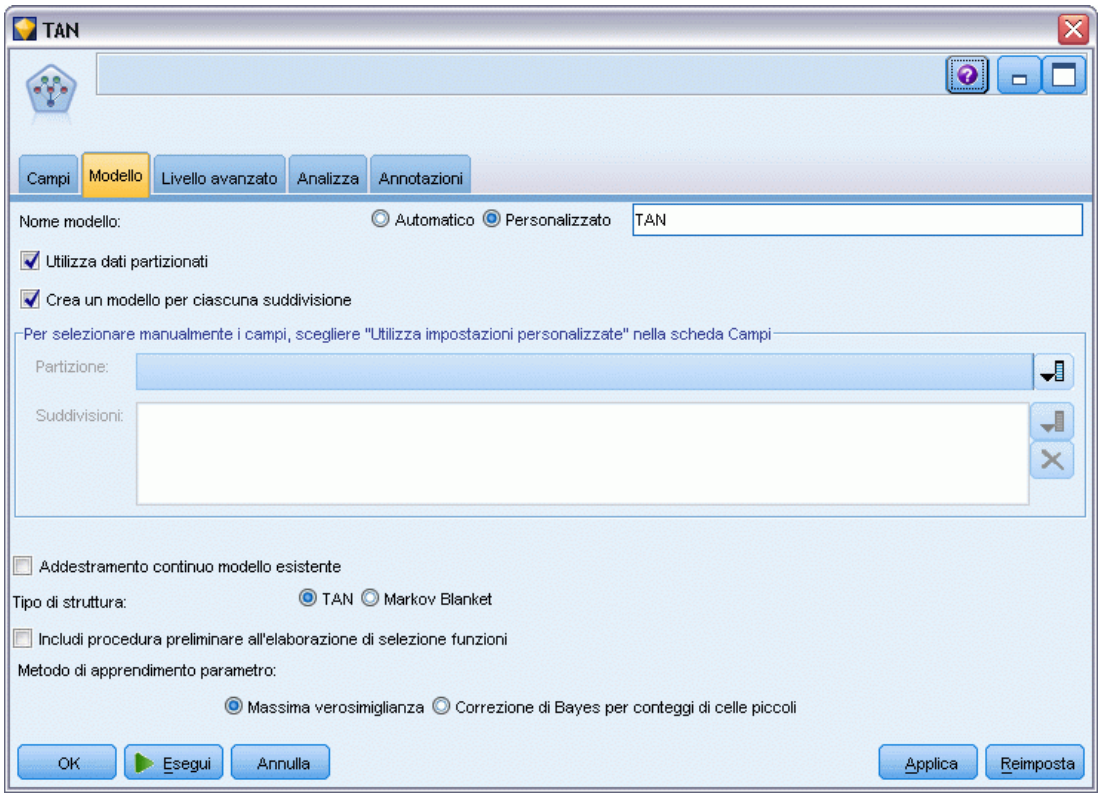
Poiché è possibile generare diversi tipi di rete bayesiana, vale la pena confrontarne alcuni per vedere quale modello fornisce le previsioni migliori. La prima rete da creare è un modello TAN (Tree Augmented Naïve Bayes).

▶ Collegare un nodo Rete bayesiana al nodo Seleziona.

▶ Nella scheda Modello, come nome del modello selezionare Personalizzato e immettere TAN nella casella di testo.

- Per il tipo Struttura, selezionare TAN e fare clic su OK.

Figura 17-4
Creazione di un modello TAN (Tree Augmented Naïve Bayes)

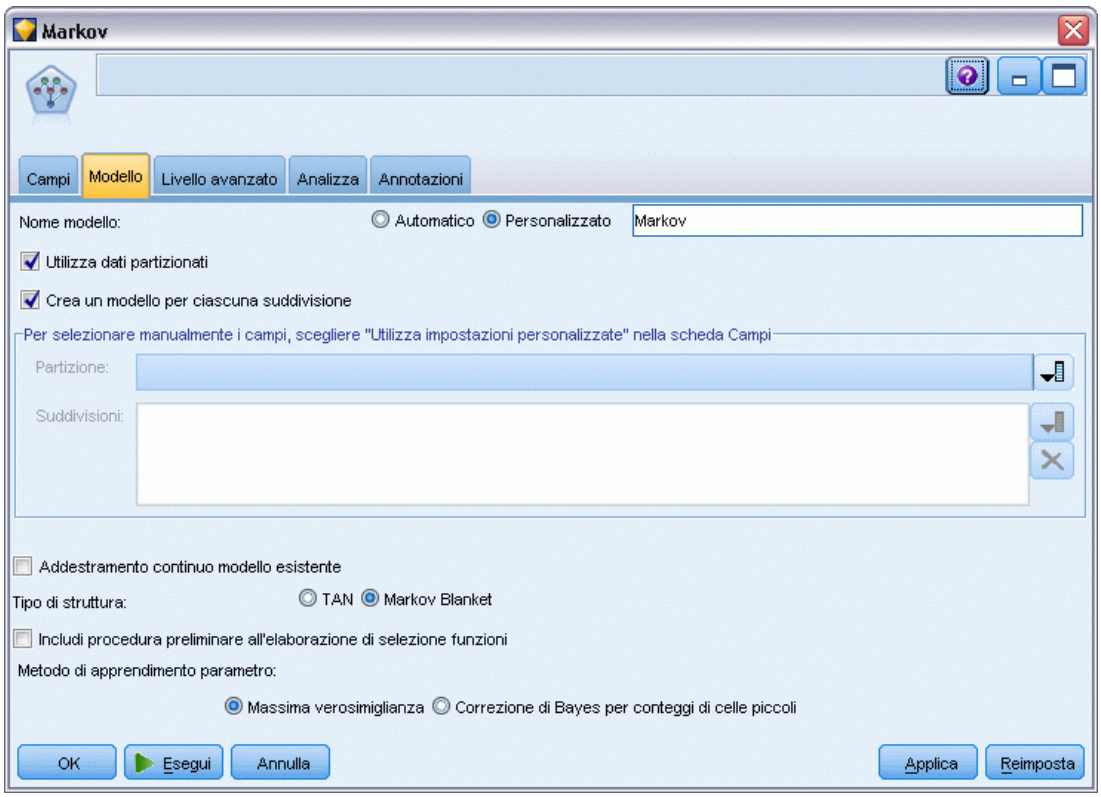


Il secondo tipo di modello da generare è una struttura Markov Blanket.

- Collegare un secondo nodo Rete bayesiana al nodo Seleziona.
- Nella scheda Modello, come nome del modello selezionare Personalizzato e immettere Markov nella casella di testo.

▶ Per il tipo Struttura, selezionare Markov Blanket e fare clic su OK.

Figura 17-5
Creazione di un modello Markov Blanket



Il terzo tipo di modello da generare è una struttura Markov Blanket. Si utilizzerà inoltre l'opzione Includi procedura preliminare all'elaborazione di selezione funzioni per selezionare gli input che sono significativamente correlati alla variabile obiettivo.

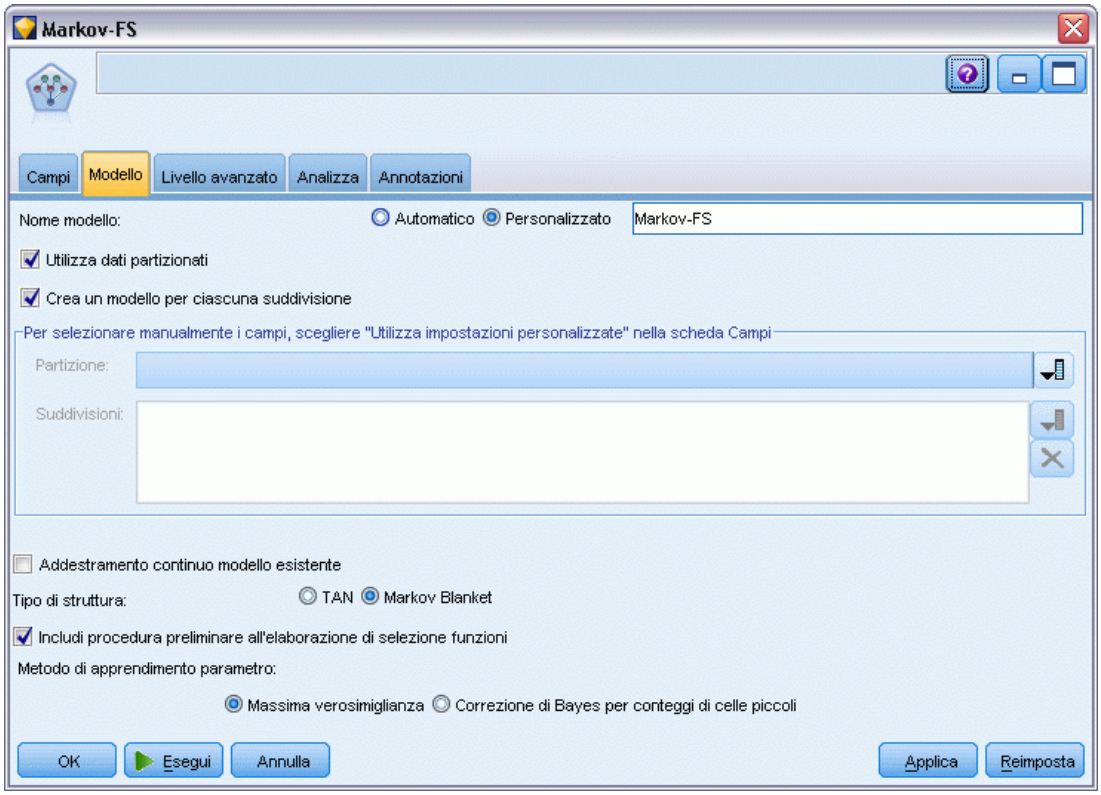
▶ Collegare un terzo nodo Rete bayesiana al nodo Seleziona.

▶ Nella scheda Modello, come nome del modello selezionare Personalizzato e immettere Markov-SF nella casella di testo.

▶ Per tipo Struttura, selezionare Markov Blanket.

▶ Selezionare Includi procedura preliminare all'elaborazione di selezione funzioni e fare clic su OK.

Figura 17-6
Creazione di un modello Markov Blanket con l'opzione Includi procedura preliminare all'elaborazione di selezione funzioni



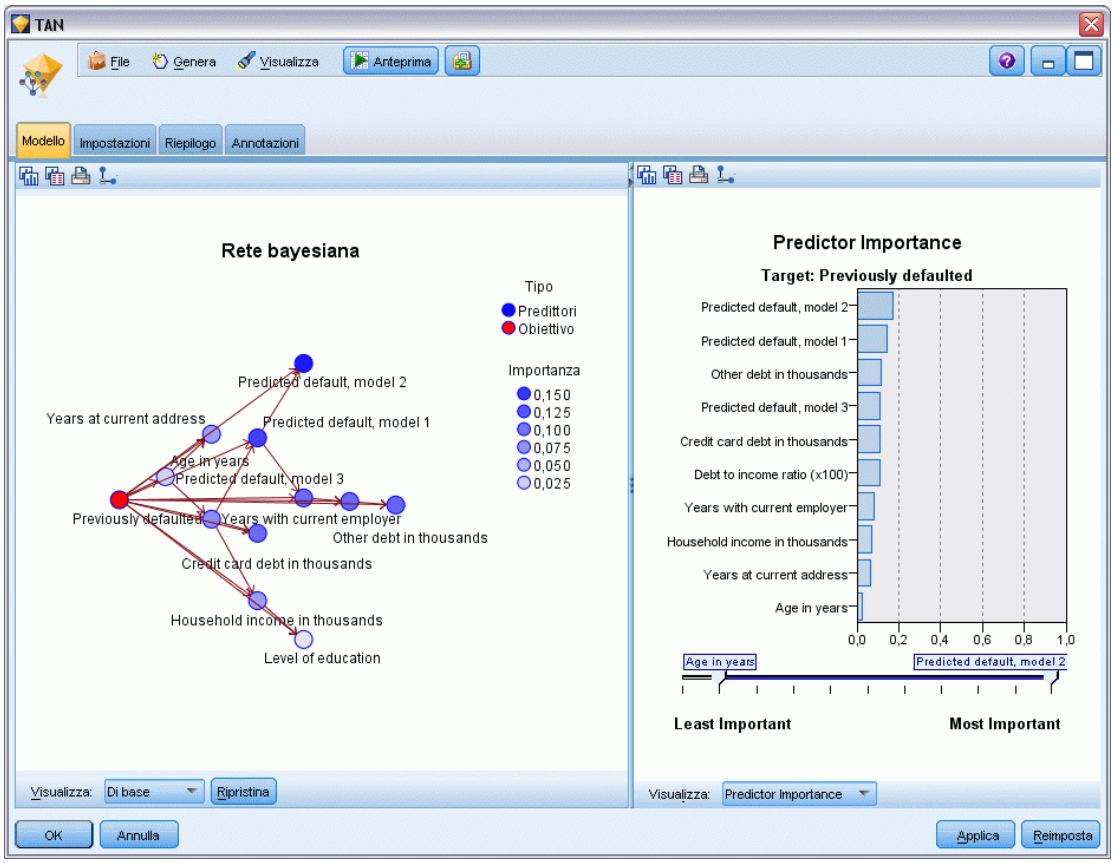
Visualizzazione del modello

▶ Eseguire lo stream per creare gli insiemi di modelli, che vengono aggiunti allo stream e alla palette Modelli nell'angolo superiore destro. Per visualizzare i rispettivi dettagli, fare doppio clic su un qualsiasi insieme di modelli dello stream.

La scheda Modello dell'insieme di modelli è suddivisa in due riquadri: il riquadro sinistro contiene una rete di nodi che mostra la relazione tra l'obiettivo e i suoi predittori più importanti, nonché la relazione tra i predittori.

Il riquadro destro mostra l'*Importanza dei predittori*, che indica l'importanza relativa di ogni predittore nella stima del modello, oppure le *Probabilità condizionali*, che contengono il valore di probabilità condizionale per ogni valore del nodo e ogni combinazione di valori nei corrispondenti nodi genitori.

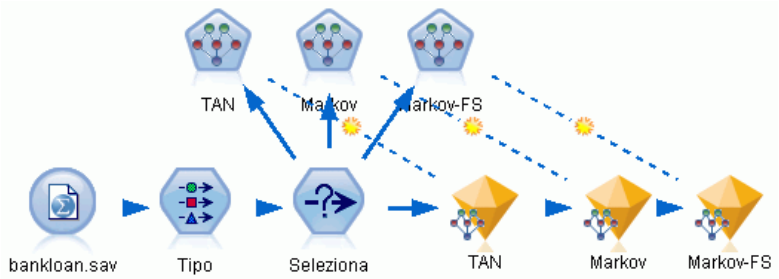
Figura 17-7
Visualizzazione di un modello TAN (Tree Augmented Naïve Bayes)



- Collegare l'insieme di modelli TAN all'insieme di modelli Markov (scegliere Sostituisci nella finestra di dialogo di avviso).
- Collegare l'insieme di modelli Markov all'insieme di modelli Markov-SF (scegliere Sostituisci nella finestra di dialogo di avviso).

▶ Per una visualizzazione più agevole, allineare i tre insiemi di modelli con il nodo Seleziona.

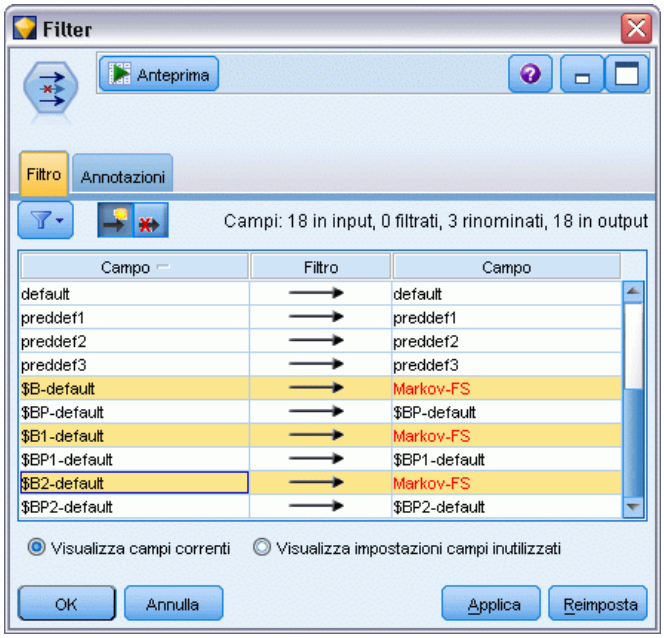
Figura 17-8
Allineamento degli insiemi di modelli nello stream



▶ Per rinominare gli output dei modelli per esigenze di chiarezza nel grafico di valutazione da creare, collegare un nodo Filtro all'insieme di modelli Markov-SF.

▶ Nella colonna *Campo* di destra, rinominare $B-default come TAN, $B1-default come Markov e $B2-default come Markov-SF.

Figura 17-9
Ridenominazione di nomi di campo del modello

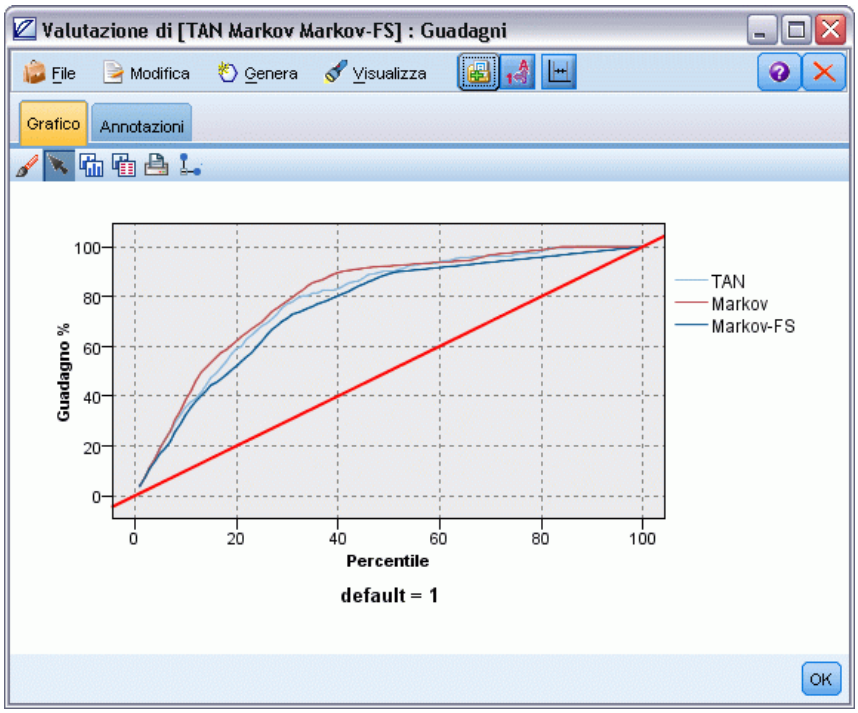


Per confrontare la precisione delle previsioni dei modelli, è possibile generare un grafico dei guadagni.

▶ Collegare un nodo Grafico di valutazione al nodo Filtro ed eseguire il nodo Grafico utilizzando le rispettive impostazioni di default.

Il grafico mostra che ogni modello produce risultati simili, tuttavia il modello Markov è leggermente migliore.

Figura 17-10
Valutazione della precisione dei modelli



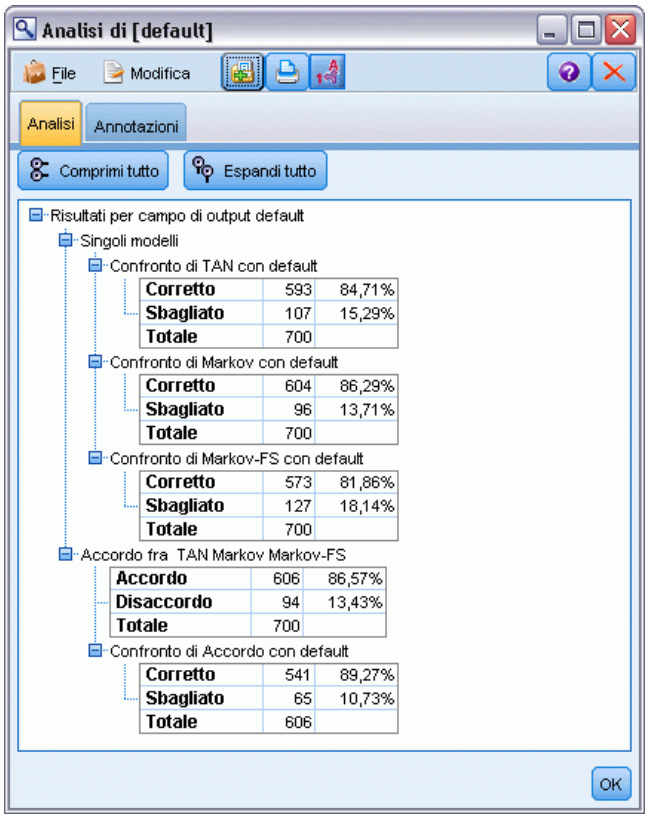
Per verificare la precisione delle previsioni di ogni modello, è possibile utilizzare un nodo Analisi anziché il Grafico di valutazione. Questo nodo mostra la precisione in termini di percentuale di previsioni corrette ed errate.

- Collegare un nodo Analisi al nodo Filtro ed eseguire il nodo Analisi utilizzando le rispettive impostazioni di default.

Analogamente a quanto avvenuto con il nodo Grafico di valutazione, anche in questo caso il modello Markov si dimostra leggermente superiore agli altri in termini di precisione delle previsioni, seguito a solo pochi punti percentuali dal modello Markov-SF. Ciò significa che potrebbe essere meglio utilizzare il modello Markov-SF, poiché questo utilizza un numero

inferiore di input per calcolare i risultati e consente pertanto di accelerare la raccolta dati e i tempi di immissione ed elaborazione.

Figura 17-11
Analisi della precisione del modello



Per una spiegazione dei fondamenti matematici dei metodi di modellazione utilizzati in IBM® SPSS® Modeler, vedere il manuale *SPSS Modeler Algorithms Guide*, disponibile nella directory *\Documentation* del disco di installazione.

Si noti inoltre che questi risultati si basano soltanto sui dati di addestramento. Per un'indicazione dell'efficacia con cui il modello potrà essere esteso ad altri dati in una situazione reale, utilizzare un nodo Partizione per estrarre un sottoinsieme di record da utilizzare per test e validazione. Per ulteriori informazioni, vedere l'argomento Nodo Partizione in il capitolo 4 in *IBM SPSS Modeler 15 Nodi di input, elaborazione e output*.

Capitolo
18

Riaddestramento di un modello su base mensile (Rete bayesiana)

Le reti bayesiane consentono 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.

In questo esempio viene utilizzato lo stream denominato *bayes_churn_retrain.str*, che fa riferimento ai file di dati *telco_Jan.sav* e *telco_Feb.sav*. Tali file sono disponibili nella directory *Demos* dell'installazione di IBM® SPSS® Modeler ed è possibile accedervi dal gruppo di programmi IBM® SPSS® Modeler nel menu Start di Windows. Il file *bayes_churn_retrain.str* si trova nella directory *streams*.

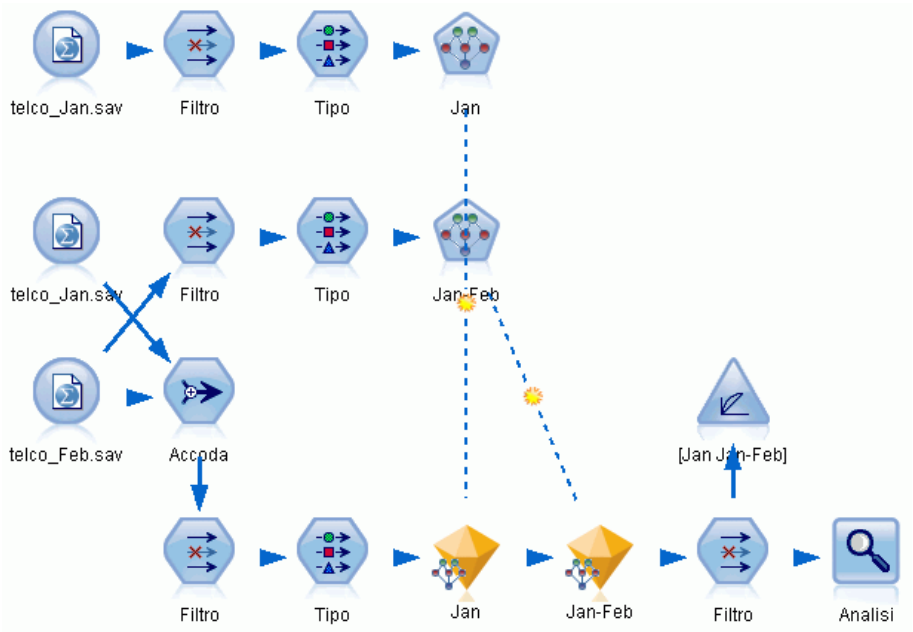
Si supponga, per esempio, che una società che fornisce servizi di telecomunicazione sia preoccupata del numero dei clienti persi a favore della concorrenza (tasso di abbandono). Potendo utilizzare i dati dei clienti per prevedere quali clienti hanno le maggiori probabilità di abbandono in futuro, quest'ultima potrà offrire loro incentivi mirati o altre offerte per scoraggiarli a lasciare l'azienda a favore di un altro fornitore di servizi.

Questo esempio utilizza i dati esistenti sul tasso di abbandono mensile per prevedere quali clienti hanno la maggiore probabilità di lasciare la società in futuro. Aggiungendovi i dati del mese successivo, è inoltre possibile ottimizzare e riaddestrare il modello.

Creazione dello stream

- Aggiungere un nodo di input File Statistics che punta a *telco_Jan.sav* nella cartella *Demos*.

Figura 18-1
Stream di esempio di rete bayesiana

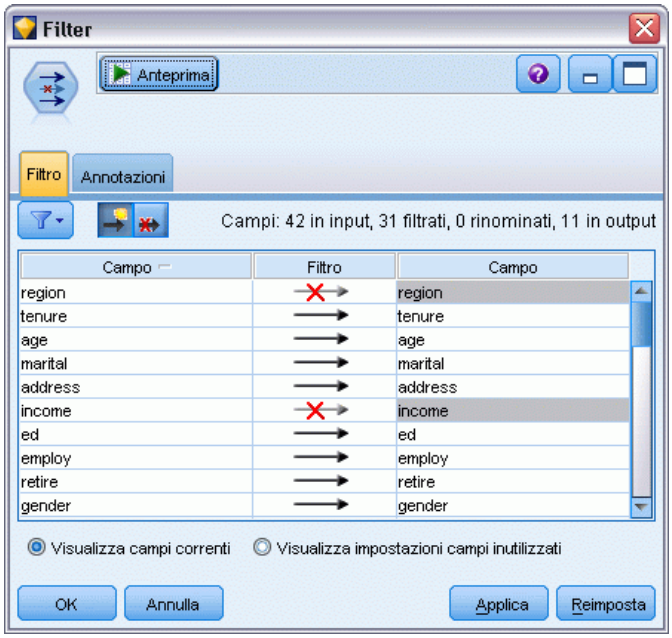


Un'analisi precedente ha rivelato che numerosi campi di dati sono ininfluenti ai fini della previsione del tasso di abbandono. È possibile filtrare tali dati dall'insieme di dati per aumentare la velocità di elaborazione quando si generano modelli e se ne calcola il punteggio.

- Aggiungere un nodo Filtro al nodo Input.
- Escludere tutti i campi, eccetto *indirizzo*, *età*, *abbandono*, *catcli*, *istruz*, *impiego*, *sesso*, *statciv*, *residenza*, *pensionato/a* e *durata*.

▶ Fare clic su OK.

Figura 18-2
Filtraggio di campi non necessari



▶ Aggiungere un nodo Tipo al nodo Filtro.

▶ Aprire il nodo Tipo e fare clic sul pulsante Leggi valori per popolare la colonna *Valori*.

- Affinché il nodo Valutazione possa stabilire quale valore è vero e quale è falso, impostare su Flag il livello di misurazione per il campo *abbandono* e il ruolo su Obiettivo. Fare clic su OK.

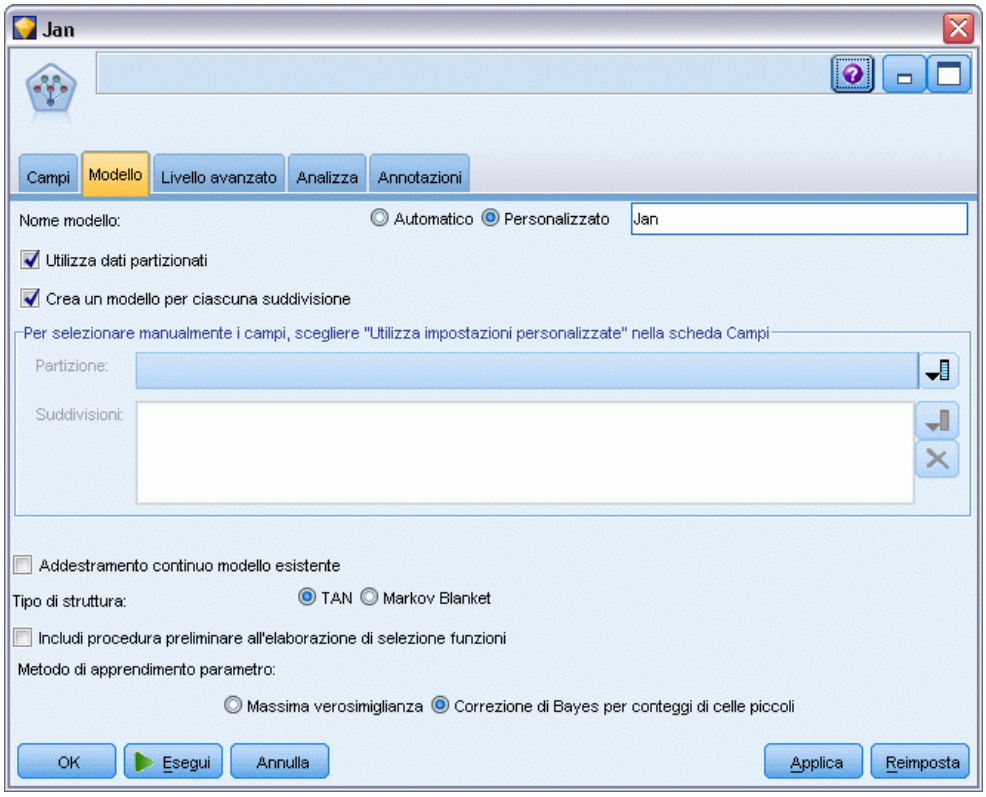
Figura 18-3
Selezione del campo obiettivo

È possibile generare diversi tipi di rete bayesiana, tuttavia ai fini di questo esempio si creerà un modello TAN (Tree Augmented Naïve Bayes). Verrà generata una rete di grandi dimensioni. Assicurarsi di includere tutti i possibili collegamenti tra le variabili dei dati per generare un robusto modello iniziale.

- Collegare un nodo Rete bayesiana al nodo Tipo.
- Nella scheda Modello, come nome del modello selezionare Personalizzato e immettere Gen nella casella di testo.
- Come Metodo di apprendimento parametro, selezionare Correzione di Bayes per conteggi di celle piccoli.

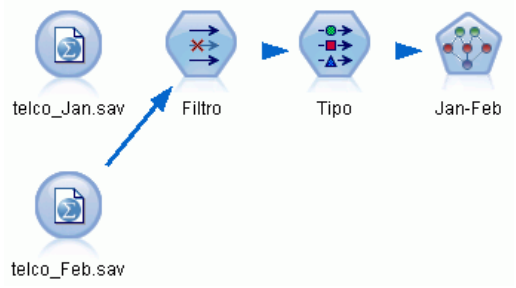
- Fare clic su Esegui. L'insieme di modelli viene aggiunto allo stream e alla palette Modelli nell'angolo superiore destro.

Figura 18-4
Creazione di un modello TAN (Tree Augmented Naïve Bayes)



- Aggiungere un nodo di input File Statistics che punta a *telco_Feb.sav* nella cartella *Demos*.
- Collegare questo nuovo nodo di input al nodo Filtro (nella finestra di dialogo di avviso, scegliere Sostituisci per sostituire la connessione al nodo di input precedente).

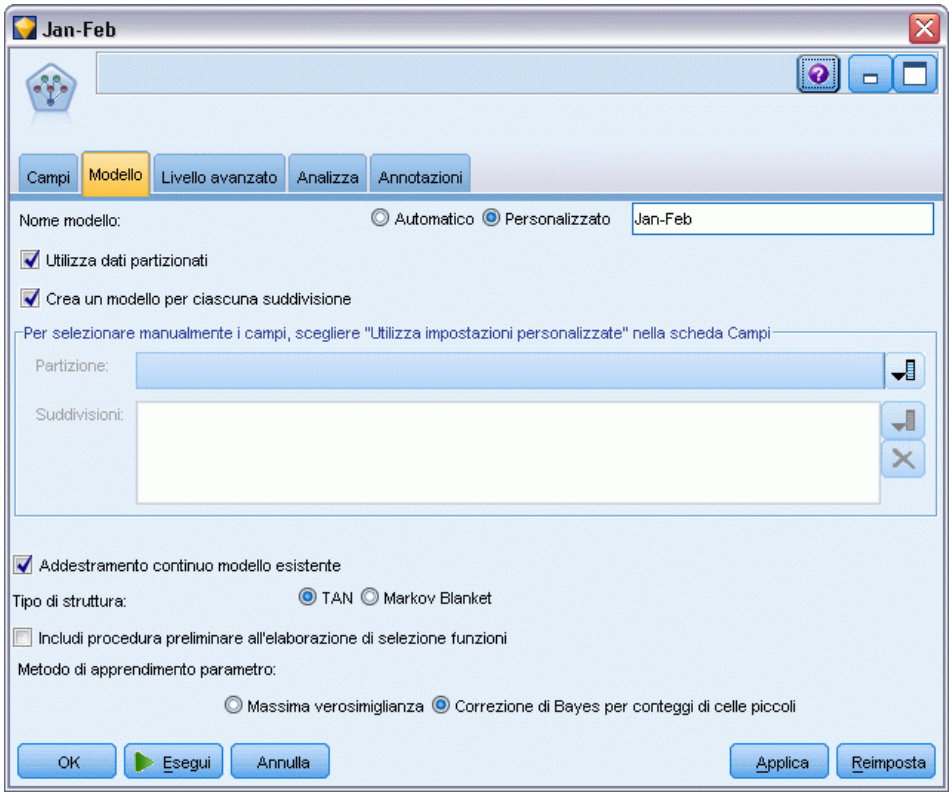
Figura 18-5
Aggiunta dei dati del secondo mese



- Nella scheda Modello del nodo Rete bayesiana, come nome del modello selezionare Personalizzato e immettere Gen-Feb nella casella di testo.
- Selezionare Addestramento continuo modello esistente.

- Fare clic su Esegui. L'insieme di modelli sovrascrive quello esistente nello stream, ma viene anche aggiunto alla palette Modelli nell'angolo superiore destro.

Figura 18-6
Riaddestramento del modello



Valutazione del modello

Per confrontare i modelli, è necessario combinare i due insiemi di dati.

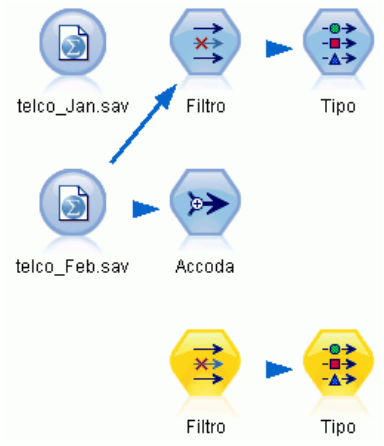
▶ Aggiungere un nodo Accoda e collegarvi i nodi di input *telco_Jan.sav* e *telco_Feb.sav*.

Figura 18-7
Collegamento delle due sorgenti dati

▶ Copiare i nodi Filtro e Tipo riportati in precedenza nello stream e incollarli nell'area di disegno dello stream.

▶ Collegare il nodo Accoda al nodo Filtro appena copiato.

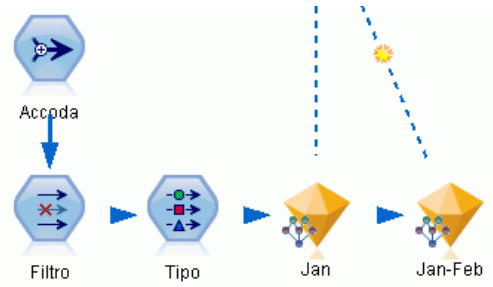
Figura 18-8
Inserimento dei nodi copiati nello stream con il comando Incolla



Gli insiemi dei due modelli Rete bayesiana si trovano nella palette Modelli nell'angolo superiore destro.

- Fare doppio clic sull'insieme di modelli Gen per aggiungerlo allo stream e collegarlo al nodo Tipo appena copiato.
- Collegare l'insieme di modelli Gen-Feb già nello stream all'insieme di modelli Gen.
- Aprire l'insieme di modelli Gen.

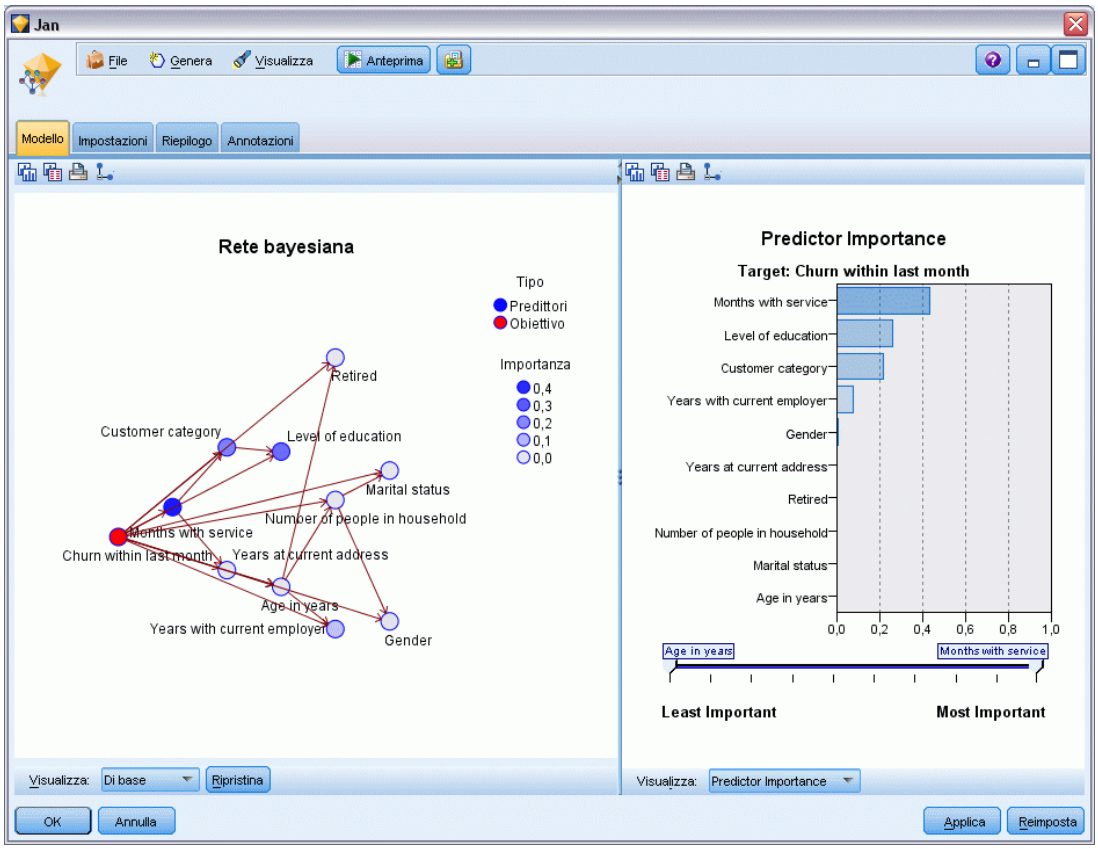
Figura 18-9
Aggiunta degli insiemi di modelli allo stream



La scheda Modello dell'insieme di modelli Rete bayesiana è suddivisa in due colonne: la colonna sinistra contiene una rete di nodi che mostra la relazione tra l'obiettivo e i suoi predittori più importanti, nonché la relazione tra i predittori.

La colonna destra mostra l'*Importanza dei predittori*, che indica l'importanza relativa di ogni predittore nella stima del modello, oppure le *Probabilità condizionali*, che contengono il valore di probabilità condizionale per ogni valore del nodo e ogni combinazione di valori nei corrispondenti nodi genitori.

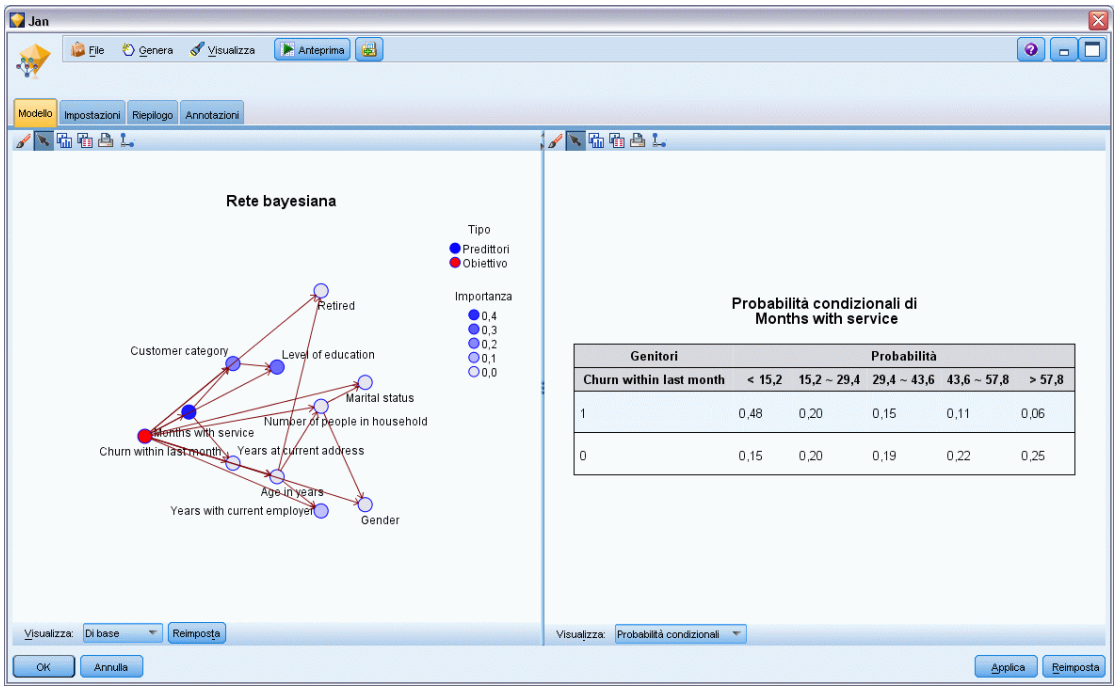
Figura 18-10
Modello Rete bayesiana che mostra l'importanza dei predittori



Per visualizzare le probabilità condizionali relative a un nodo, fare clic sul nodo desiderato nella colonna sinistra. La colonna destra viene aggiornata per mostrare le informazioni richieste.

Vengono visualizzate le probabilità condizionali per ogni bin in cui sono stati divisi i valori di dati rispetto al genitore del nodo e ai nodi di pari livello.

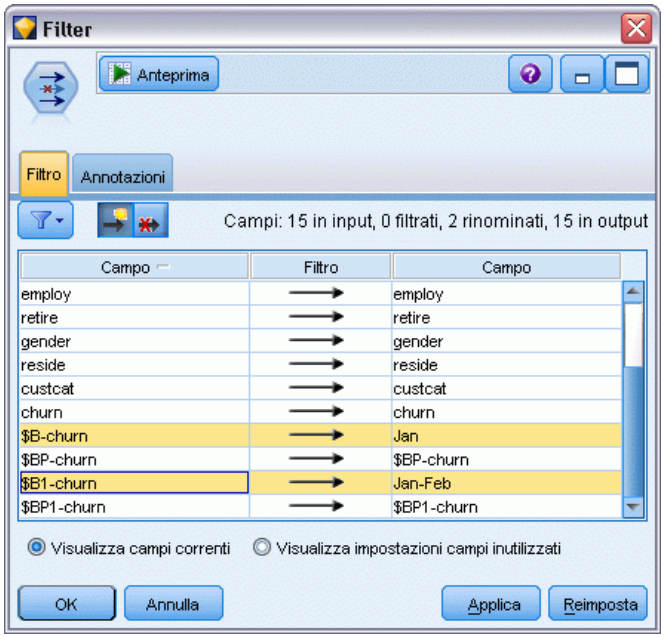
Figura 18-11
Modello Rete bayesiana che mostra le probabilità condizionali



- Per rinominare gli output del modello per esigenze di chiarezza, collegare un nodo Filtro all'insieme di modelli Gen-Feb.

▶ Nella colonna *Campo* di destra, rinominare $B-churn come Gen e $B1-churn come Gen-Feb.

Figura 18-12
Ridenominazione di nomi di campo del modello



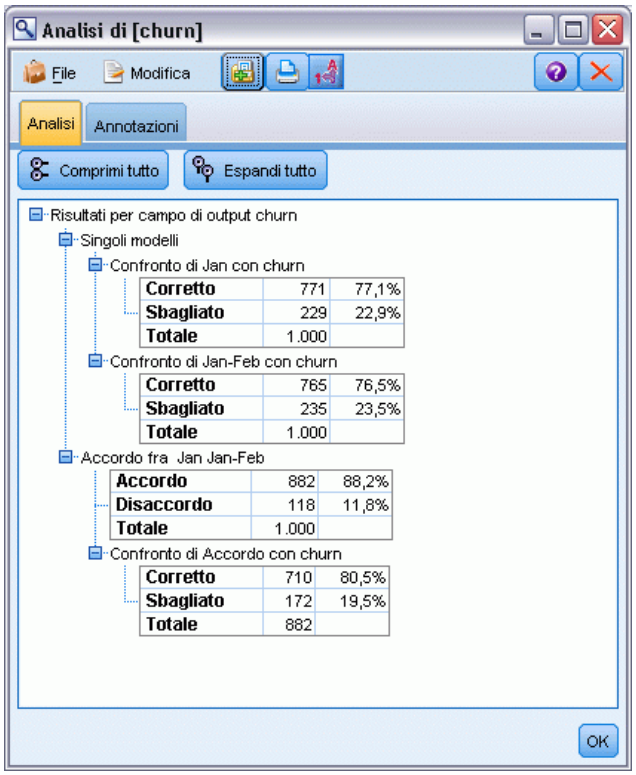
Per verificare la precisione delle previsioni dei modelli, è possibile utilizzare un nodo Analisi. Questo nodo mostra la precisione in termini di percentuale di previsioni corrette ed errate.

▶ Collegare un nodo Analisi al nodo Filtro.

▶ Aprire il nodo Analisi e fare clic su Esegui.

In questo modo si evidenzia che entrambi i modelli possiedono un grado di precisione simile nella previsione del tasso di abbandono.

Figura 18-13
Analisi della precisione del modello



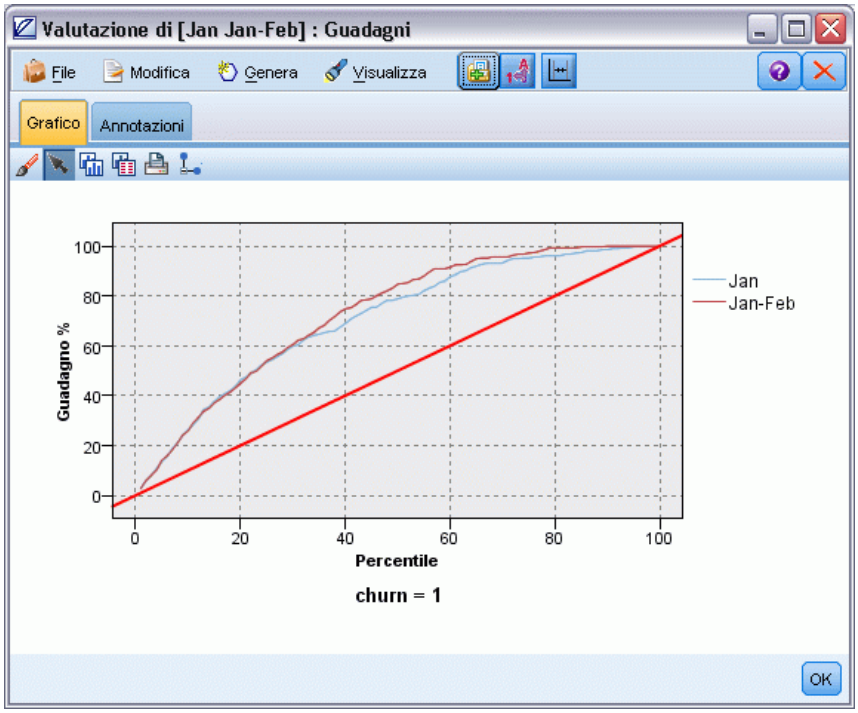
In alternativa al nodo Analisi, è possibile utilizzare un grafico di valutazione per confrontare la precisione delle previsioni dei modelli generando un grafico dei guadagni.

- Collegare un nodo grafico Valutazione al nodo Filtro.

Eseguire il nodo Grafico utilizzando le sue impostazioni di default.

Analogamente a quanto si è verificato con il nodo Analisi, il grafico mostra che ogni tipo di modello produce risultati simili, tuttavia il modello riaddestrato, che utilizza i dati di entrambi i mesi, è leggermente migliore perché dispone di un maggior livello di confidenza nelle sue previsioni.

Figura 18-14
Valutazione della precisione dei modelli



Per una spiegazione dei fondamenti matematici dei metodi di modellazione utilizzati in IBM® SPSS® Modeler, vedere il manuale *SPSS Modeler Algorithms Guide*, disponibile nella directory *\Documentation* del disco di installazione.

Si noti inoltre che questi risultati si basano soltanto sui dati di addestramento. Per un'indicazione dell'efficacia con cui il modello potrà essere esteso ad altri dati in una situazione reale, utilizzare un nodo Partizione per estrarre un sottoinsieme di record da utilizzare per test e validazione. Per ulteriori informazioni, vedere l'argomento Nodo Partizione in il capitolo 4 in *IBM SPSS Modeler 15 Nodi di input, elaborazione e output*.

Capitolo
19

Promozione nella vendita al dettaglio (rete neurale/C&RT)

In questo esempio vengono utilizzati i dati relativi a linee di prodotti per la vendita al dettaglio e vengono illustrati gli effetti delle promozioni sulle vendite. (I dati sono fittizi). L'obiettivo è prevedere gli effetti delle promozioni future sulle vendite. Come nell'esempio relativo al monitoraggio della condizione, il processo di data mining prevede le fasi di analisi, preparazione dei dati, addestramento e verifica.

In questo esempio vengono utilizzati gli stream denominati *goodsplot.str* e *goodslearn.str*, che fanno riferimento ai file di dati denominati *GOODS1n* e *GOODS2n*. Questi file sono disponibili nella directory *Demos* di tutte le installazioni di IBM® SPSS® Modeler. L'accesso si effettua dal gruppo di programmi IBM® SPSS® Modeler nel menu Start di Windows. Lo stream *goodsplot.str* si trova nella cartella *streams* e il file *goodslearn.str* nella directory *streams*.

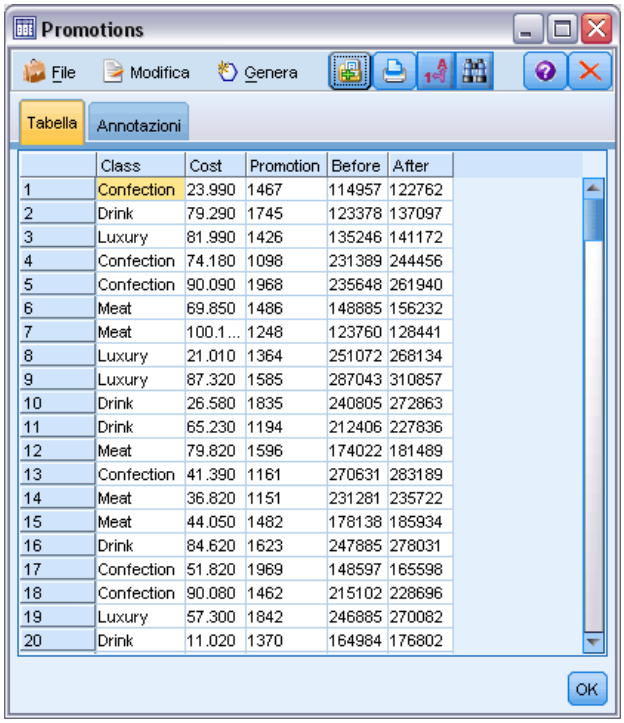
Esame dei dati

Ogni record contiene:

- *Categoria*. Tipo di prodotto.
- *Costo*. Prezzo unitario.
- *Promozione*. Indice dell'importo speso per una promozione specifica.
- *Prima*. Entrata prima della promozione.
- *Dopo*. Entrata dopo la promozione.

Lo stream *goodsplot.str* contiene uno stream semplice per la visualizzazione dei dati in una tabella. I due campi relativi all'entrata (*Prima* e *Dopo*) sono espressi in termini assoluti. È tuttavia probabile che l'incremento dell'entrata dopo la promozione (e presumibilmente in seguito ad essa) possa essere maggiormente significativo.

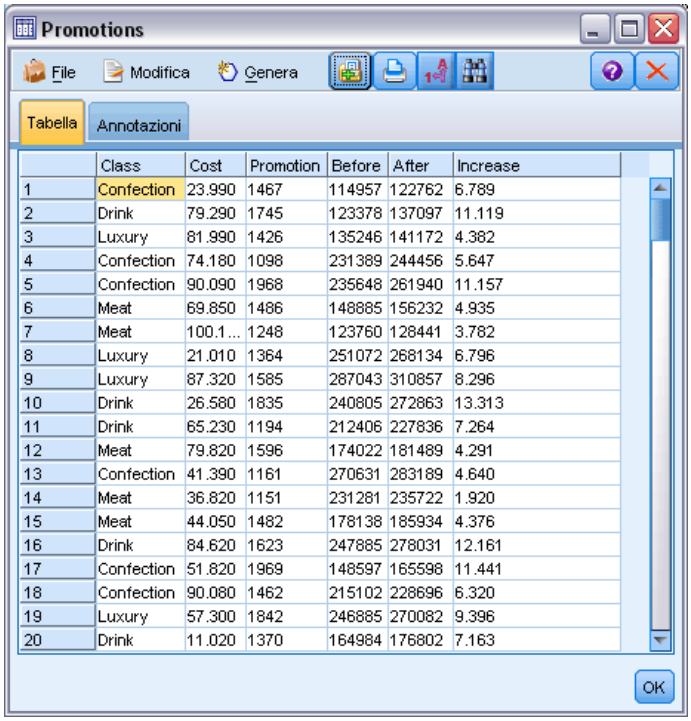
Figura 19-1
Effetti della promozione sulle vendite dei prodotti



	Class	Cost	Promotion	Before	After
1	Confection	23.990	1467	114957	122762
2	Drink	79.290	1745	123378	137097
3	Luxury	81.990	1426	135246	141172
4	Confection	74.180	1098	231389	244456
5	Confection	90.090	1968	235648	261940
6	Meat	69.850	1486	148885	156232
7	Meat	100.1...	1248	123760	128441
8	Luxury	21.010	1364	251072	268134
9	Luxury	87.320	1585	287043	310857
10	Drink	26.580	1835	240805	272863
11	Drink	65.230	1194	212406	227836
12	Meat	79.820	1596	174022	181489
13	Confection	41.390	1161	270631	283189
14	Meat	36.820	1151	231281	235722
15	Meat	44.050	1482	178138	185934
16	Drink	84.620	1623	247885	278031
17	Confection	51.820	1969	148597	165598
18	Confection	90.080	1462	215102	228696
19	Luxury	57.300	1842	246885	270082
20	Drink	11.020	1370	164984	176802

goodsplot.str contiene anche un nodo per derivare questo valore, espresso come percentuale dell'entrata prima della promozione, in un campo denominato *Incremento* e visualizza una tabella che include il campo.

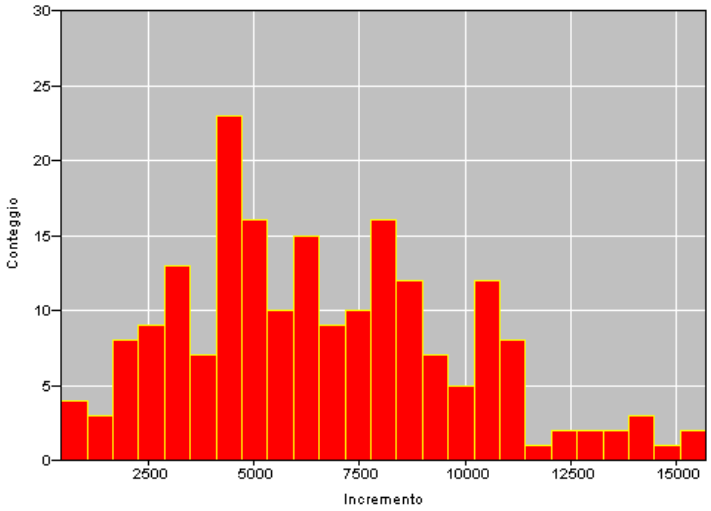
Figura 19-2
Incremento dell'entrata dopo la promozione



	Class	Cost	Promotion	Before	After	Increase
1	Confection	23.990	1467	114957	122762	6.789
2	Drink	79.290	1745	123378	137097	11.119
3	Luxury	81.990	1426	135246	141172	4.382
4	Confection	74.180	1098	231389	244456	5.647
5	Confection	90.090	1968	235648	261940	11.157
6	Meat	69.850	1486	148885	156232	4.935
7	Meat	100.1...	1248	123760	128441	3.782
8	Luxury	21.010	1364	251072	268134	6.796
9	Luxury	87.320	1585	287043	310857	8.296
10	Drink	26.580	1835	240805	272863	13.313
11	Drink	65.230	1194	212406	227836	7.264
12	Meat	79.820	1596	174022	181489	4.291
13	Confection	41.390	1161	270631	283189	4.640
14	Meat	36.820	1151	231281	235722	1.920
15	Meat	44.050	1482	178138	185934	4.376
16	Drink	84.620	1623	247885	278031	12.161
17	Confection	51.820	1969	148597	165598	11.441
18	Confection	90.080	1462	215102	228696	6.320
19	Luxury	57.300	1842	246885	270082	9.396
20	Drink	11.020	1370	164984	176802	7.163

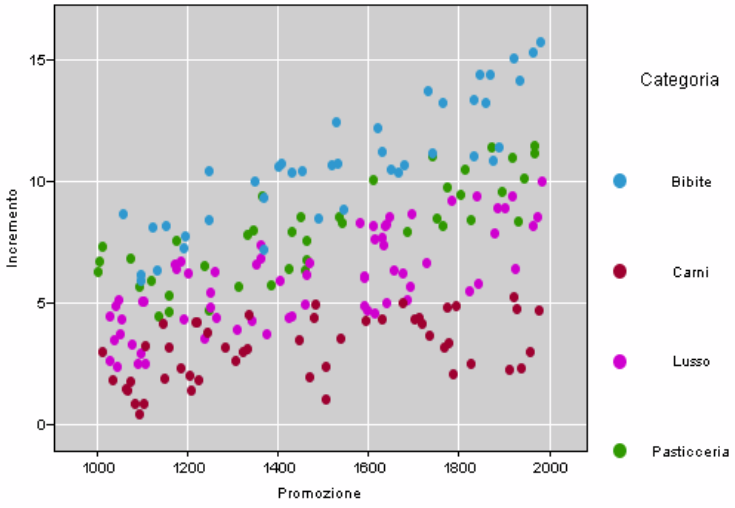
Lo stream visualizza inoltre un istogramma dell'incremento e un grafico a dispersione che rappresenta l'incremento in relazione ai costi della promozione, a cui è sovrapposta la categoria del prodotto specifico.

Figura 19-3
Istogramma dell'incremento dell'entrata



Dal grafico a dispersione risulta che per ogni classe di prodotto esiste una relazione quasi lineare tra l'incremento dell'entrata e i costi della promozione. È quindi probabile che un albero decisionale o una rete neurale siano in grado di prevedere l'incremento dell'entrata dagli altri campi disponibili con una precisione adeguata.

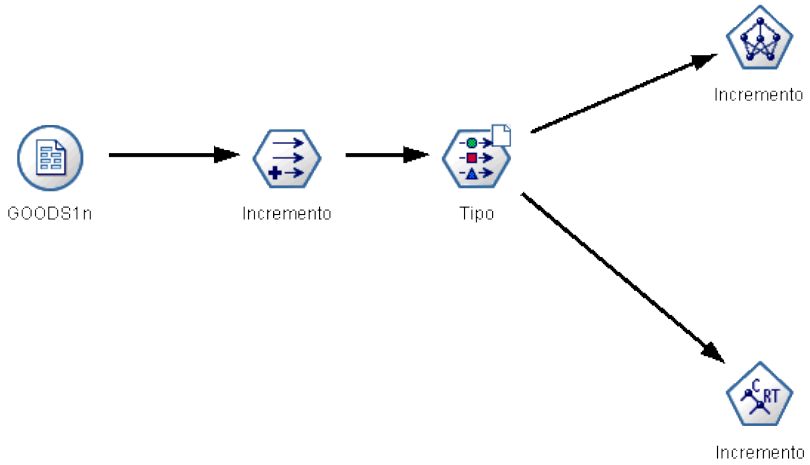
Figura 19-4
Incremento dell'entrata in relazione alle spese promozionali



Apprendimento e verifica

Lo stream *goodslearn.str* addestra una rete neurale e un albero decisionale ad effettuare questa previsione dell'incremento dell'entrata.

Figura 19-5
Modeling dello stream goodslearn.str



Dopo avere eseguito i nodi del modello e generato i modelli effettivi, è possibile verificare i risultati del processo di apprendimento. A tale scopo, connettere l'albero decisionale e la rete in serie tra il nodo Tipo e un nuovo nodo Analisi, modificando il file di input dei dati in *GOODS2n* ed eseguendo il nodo Analisi. Dall'output di questo nodo e in particolare dalla correlazione lineare tra l'incremento previsto e la risposta corretta, risulterà che i sistemi addestrati prevedono l'incremento dell'entrata con una maggiore probabilità di successo.

Eventuali analisi aggiuntive potrebbero essere incentrate sui casi per i quali i sistemi addestrati generano errori di entità piuttosto considerevole e che possono essere identificati creando un grafico in cui sono rappresentati l'incremento dell'entrata previsto in relazione a quello effettivo. I valori anomali del grafico possono essere selezionati utilizzando i grafici interattivi di IBM® SPSS® Modeler. Nelle relative proprietà è possibile ottimizzare la descrizione dei dati o il processo di apprendimento, in modo da aumentarne la precisione.

Capitolo
20

Monitoraggio della condizione (Rete neurale/C5.0)

In questo esempio verranno monitorate le informazioni sullo stato da un computer e verrà illustrato il problema dell'individuazione e della previsione degli stati di errore. I dati vengono creati mediante una simulazione fittizia e consistono in diverse serie concatenate misurate nel corso del tempo. Ogni record è un report tipo "istantanea" relativo a quanto segue:

- *Tempo*. Un intero.
- *Potenza*. Un intero.
- *Temperatura*. Un intero.
- *Pressione*. 0 se è normale, 1 per un avviso temporaneo relativo alla pressione.
- *Tempo attività*. Tempo trascorso dall'ultimo servizio.
- *Stato*. In condizione normale è 0, ma in caso di errore viene sostituito da un codice di errore (101, 202 o 303).
- *Risultato*. Il codice di errore visualizzato nella serie storica oppure 0 se non si verifica alcun errore. (Questi codici sono disponibili solo al termine dell'esecuzione).

In questo esempio vengono utilizzati gli stream denominati *condplot.str* e *condlearn.str*, che fanno riferimento ai file di dati denominati *COND1n* e *COND2n*. Questi file sono disponibili nella directory *Demos* di tutte le installazioni di IBM® SPSS® Modeler. L'accesso si effettua dal gruppo di programmi IBM® SPSS® Modeler nel menu Start di Windows. I file *condplot.str* e *condlearn.str* si trovano nella directory *streams*.

Per ogni serie storica, è disponibile una serie di record relativa a un periodo di funzionamento normale, seguito da un periodo con la condizione di errore, come illustrato nella tabella seguente:

Time	Potenza	Temperatura	Pressione	Tempo attività	Stato	Risultato
0	1059	259	0	404	0	0
1	1059	259	0	404	0	0
			...			
51	1059	259	0	404	0	0
52	1059	259	0	404	0	0
53	1007	259	0	404	0	303
54	998	259	0	404	0	303
			...			
89	839	259	0	404	0	303
90	834	259	0	404	303	303
0	965	251	0	209	0	0
1	965	251	0	209	0	0
			...			

Time	Potenza	Temperatura	Pressione	Tempo attività	Stato	Risultato
51	965	251	0	209	0	0
52	965	251	0	209	0	0
53	938	251	0	209	0	101
54	936	251	0	209	0	101
			...			
208	644	251	0	209	0	101
209	640	251	0	209	101	101

Nella maggior parte dei progetti di data mining viene eseguito il seguente processo:

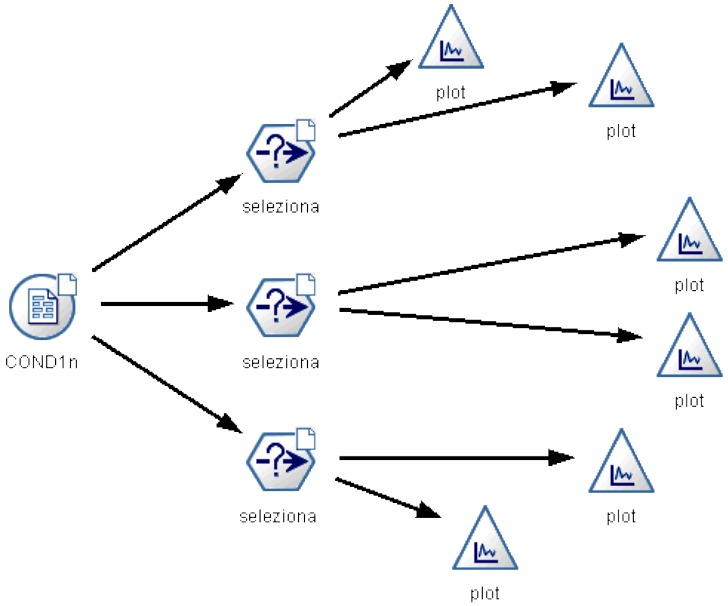
- Esame dei dati allo scopo di determinare gli attributi che possono essere rilevanti per la previsione o l'individuazione degli stati desiderati.
- Conservazione degli attributi (se già presenti) oppure derivazione e aggiunta degli attributi ai dati, se necessario.
- Utilizzo dei dati risultanti per l'addestramento delle regole e delle reti neurali.
- Verifica dei sistemi addestrati mediante i dati di test indipendenti.

Esame dei dati

Il file *condplot.str* illustra la prima parte del processo. Contiene uno stream che consente di creare una serie di grafici. Se la serie storica relativa alla temperatura o alla potenza include schemi visibili, è possibile differenziare le condizioni di errore imminenti o prevederne il verificarsi. Lo stream riportato di seguito traccia le serie storiche associate ai tre diversi codici di errore sia per

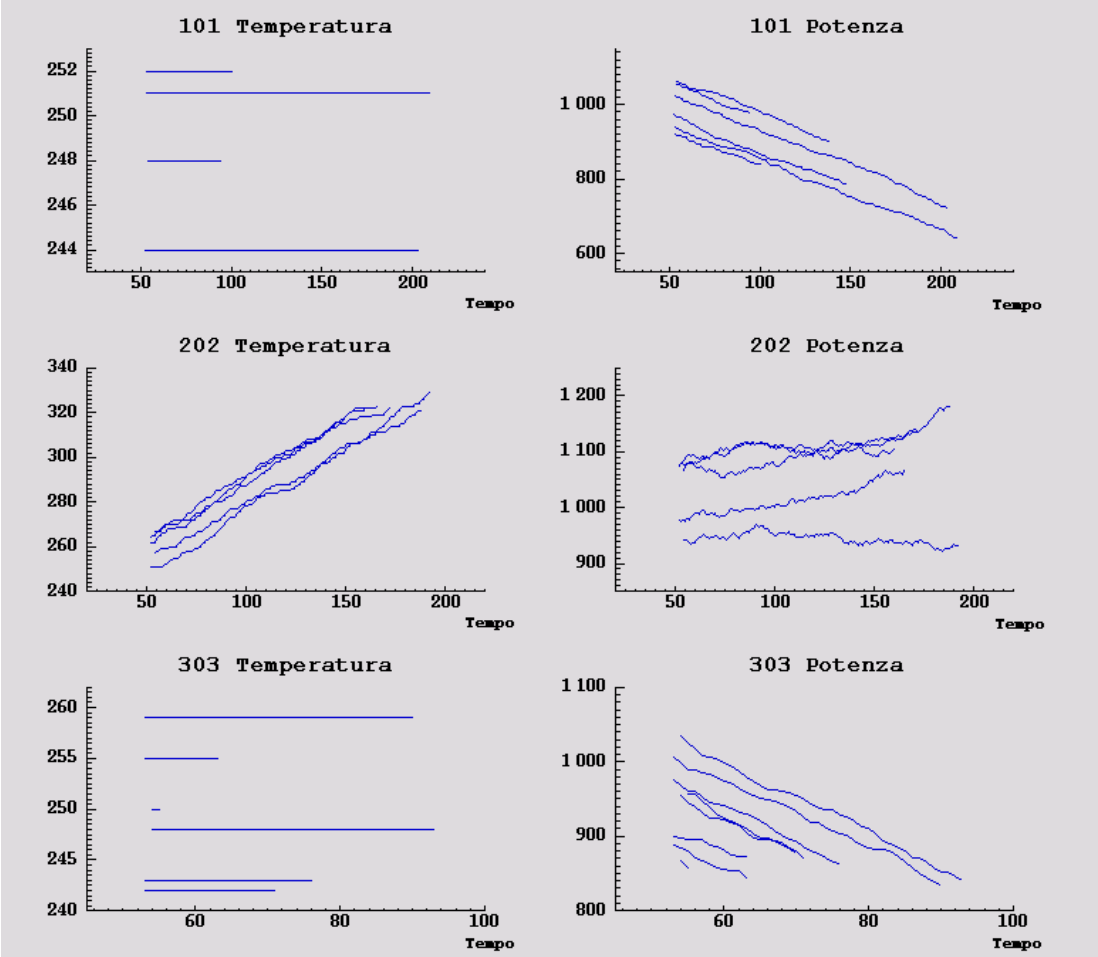
la temperatura che per la potenza e pertanto crea sei grafici. I nodi Seleziona separano i dati associati ai diversi codici di errore.

Figura 20-1
Stream condplot



I risultati dello stream sono illustrati in questa figura.

Figura 20-2
Temperatura e potenza nel tempo



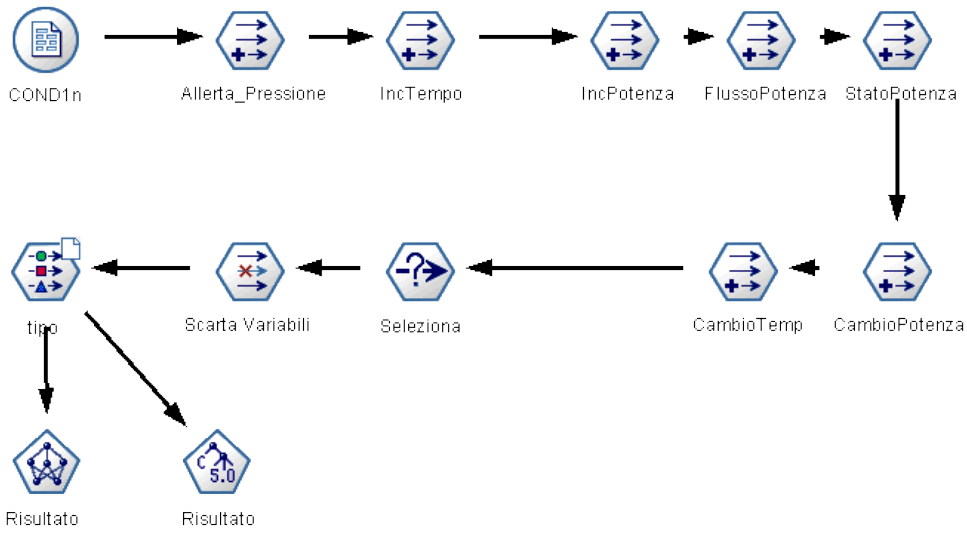
Nei grafici vengono visualizzati chiaramente gli schemi che distinguono gli errori 202 dagli errori 101 e 303. A differenza degli altri errori, gli errori 202 mostrano un aumento della temperatura e una fluttuazione della potenza nel tempo. Gli schemi che differenziano gli errori 101 dagli errori 303 sono invece meno chiari. Entrambi gli errori indicano una temperatura uniforme e una diminuzione della potenza, che tuttavia sembra più accentuata per gli errori 303.

Dai grafici risulta che la presenza e l'ampiezza della variazione sia della temperatura sia della potenza, nonché la presenza e il grado della fluttuazione, sono significativi per la previsione e l'individuazione degli errori. È pertanto consigliabile aggiungere questi attributi ai dati prima dell'applicazione dei sistemi di apprendimento.

Data Preparation

In base ai risultati dell'analisi dei dati, lo stream *condlearn.str* deriva i dati rilevanti e apprende la previsione degli errori.

Figura 20-3
Stream condlearn



Lo stream utilizza una serie di nodi Nuovo campo per preparare i dati per la modellazione.

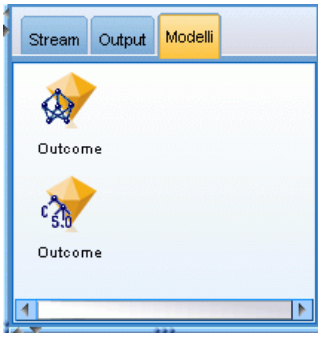
- **Nodo Testo File**. Legge il file di dati *COND1n*.
- **Nodo Nuovo campo Allerta_Pressione**. Esegue il conteggio del numero degli avvisi temporanei relativi alla pressione. Reimpostare quando il tempo torna a 0.
- **Nodo Nuovo campo IncTempo**. Calcola la variazione temporanea della temperatura utilizzando @DIFF1.
- **Nodo Nuovo campo IncPotenza**. Calcola la variazione temporanea della potenza utilizzando @DIFF1.
- **Nodo Nuovo campo FlussoPotenza**. Flag corrispondente a vero se si è verificata una variazione di potenza in direzioni opposte nell'ultimo record e in questo record, per esempio dovuta a un picco o a un passaggio di potenza.
- **Nodo Nuovo campo StatoPotenza**. Stato che inizia come *Stabile* e passa a *Fluttuante* quando vengono rilevati due flussi di potenza successivi. Torna a *Stabile* solo se non viene rilevato un flusso di potenza per cinque intervalli di tempo o se *Tempo* viene reimpostato.
- **CambioPotenza**. Media di *IncPotenza* relativa agli ultimi cinque intervalli di tempo.
- **CambioTemp**. Media di *IncTemp* relativa agli ultimi cinque intervalli di tempo.
- **Scarta Iniziali** . Scarta il primo record di ogni serie storica per evitare salti di grande entità (non corretti) in *Potenza* e *Temperatura* in prossimità dei limiti.

- **Scarta Campi**. Taglia i record a *Tempo attività*, *Stato*, *Risultato*, *Allerta_Pressione*, *StatoPotenza*, *CambioPotenza* e *CambioTemp*.
- **Tipo**. Definisce il ruolo *Obiettivo* come Risultato (il campo per cui si esegue la previsione). Definisce inoltre il livello di misurazione di *Risultato* come Nominale, *Allerta_Pressione* come Continuo e *StatoPotenza* come Flag.

Ciclo di apprendimento

L'esecuzione dello stream in *condlearn.str* consente di addestrare la regola C5.0 e la rete neurale. L'addestramento della rete può richiedere tempo, ma l'interruzione della procedura nelle fasi iniziali consente di salvare una rete in grado di produrre risultati accettabili. Al termine dell'apprendimento, la scheda Modelli nella parte in alto a destra della finestra Manager lampeggia per avvisare che sono stati creati due nuovi insiemi di modelli: uno rappresenta la rete neurale e l'altro rappresenta la regola.

Figura 20-4
Scheda Modelli della finestra Manager con gli insiemi di modelli



È anche possibile aggiungere gli insiemi di modelli allo stream esistente per verificare il sistema o esportare i risultati del modello. In questo esempio verranno verificati i risultati del modello.

Verifica

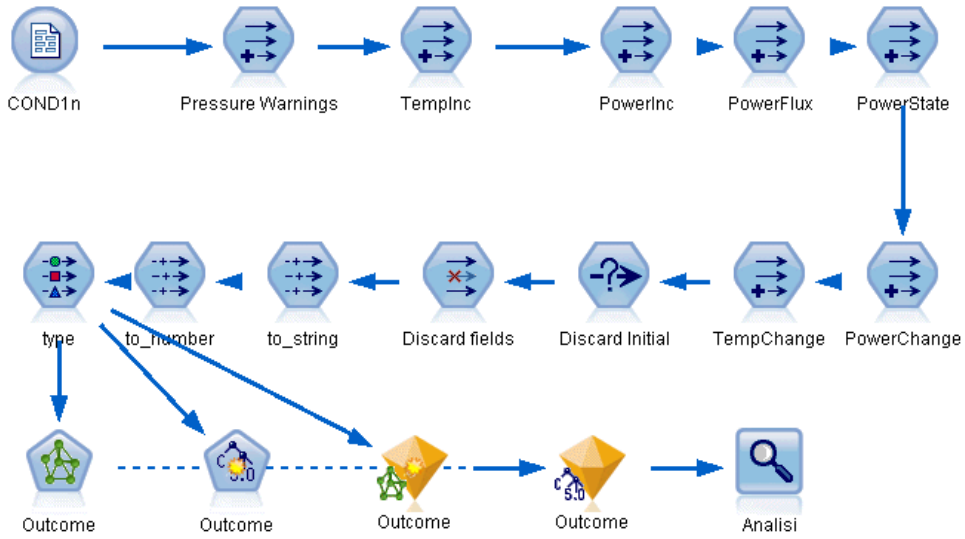
Gli insiemi di modelli vengono aggiunti allo stream, entrambi connessi al nodo Tipo.

▶ Riposizionare gli insiemi di modelli come indicato, in modo che il nodo Tipo si connetta all'insieme di modelli rete neurale, che si connette all'insieme di modelli C5.0.

▶ Collegare un nodo Analisi all'insieme di modelli C5.0.

- Modificare il nodo di input originale in modo da leggere il file *COND2n* anziché *COND1n*, in quanto *COND2n* contiene i dati del test non visibili.

Figura 20-5
Verifica della rete addestrata



- Aprire il nodo Analisi e fare clic su Esegui.

  In questo modo si ottengono valori che riflettono la precisione della regola e della rete addestrata.

Capitolo
21

Classificazione dei clienti nelle telecomunicazioni (Analisi discriminante)

L'analisi discriminante, una tecnica statistica che consente di classificare i record in base ai valori dei campi di input, È analoga alla regressione lineare ma, al posto di un campo numerico, prende un campo obiettivo categoriale.

Si supponga, per esempio, che una società che fornisce servizi di telecomunicazioni abbia segmentato la propria clientela in base a schemi di utilizzo del servizio, suddividendo i clienti in quattro gruppi. Se si utilizzano i dati demografici per prevedere l'appartenenza al gruppo, è possibile personalizzare le offerte per potenziali clienti individuali.

In questo esempio viene utilizzato lo stream denominato *telco_custcat_discriminant.str*, che fa riferimento al file di dati denominato *telco.sav*. Questi file sono disponibili nella directory *Demos* di tutte le installazioni di IBM® SPSS® Modeler. L'accesso si effettua dal gruppo di programmi IBM® SPSS® Modeler nel menu Start di Windows. Il file *telco_custcat_discriminant.str* si trova nella directory *streams*.

Questo esempio si concentra sull'utilizzo dei dati demografici per prevedere gli schemi di utilizzo dei servizi. Il campo obiettivo *catcli* ha quattro valori possibili, che corrispondono ai quattro gruppi di clienti, come segue:

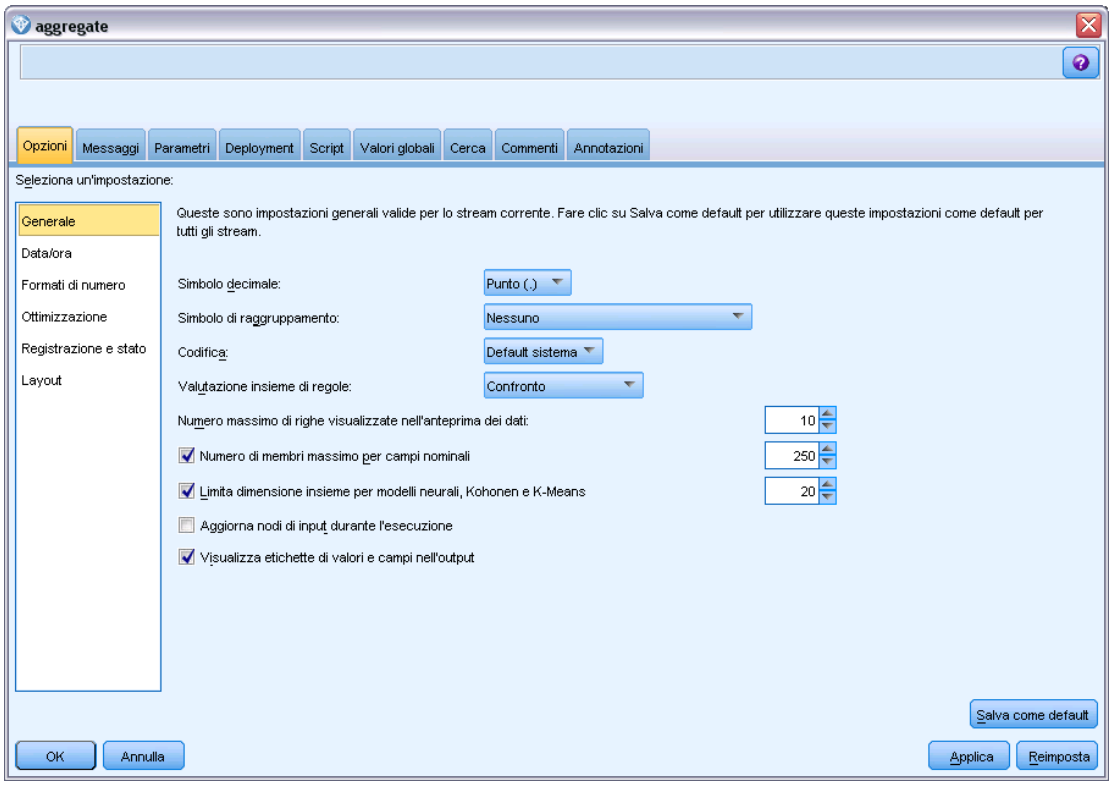
Valore	Label
1	Servizio Base
2	E-Service
3	Servizio Plus
4	Servizio Totale

Creazione dello stream

- Come prima cosa, impostare le proprietà dello stream in modo da visualizzare le etichette di variabili e valori nell'output. Dai menu, scegliere:
  File > Proprietà stream... > Opzioni > Generale

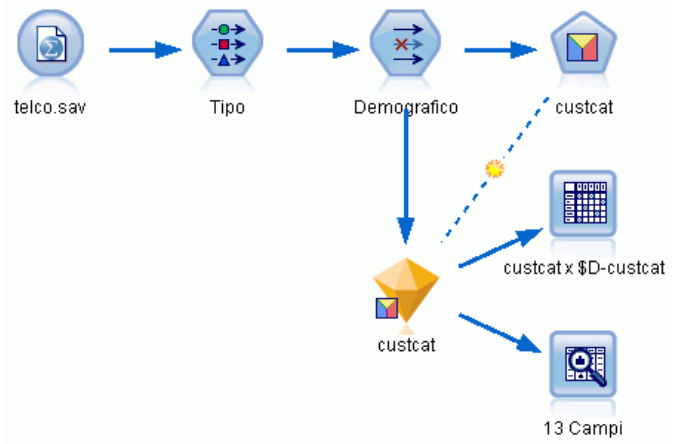
- Assicurarsi che l'opzione Visualizza etichette di valori e campi nell'output sia selezionata e fare clic su OK.

Figura 21-1
Proprietà degli stream



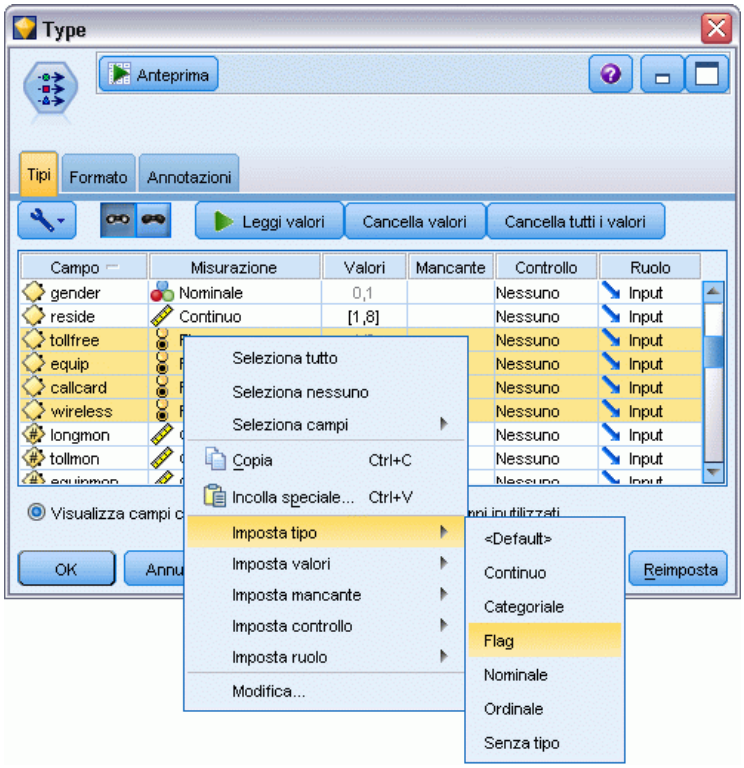
- Aggiungere un nodo di input File Statistics che punta a *telco.sav* nella cartella *Demos*.

Figura 21-2
Stream campione per la classificazione dei clienti mediante l'analisi discriminante



- Aggiungere un nodo Tipo e fare clic su Leggi valori, assicurandosi che tutti i livelli di misurazione siano impostati correttamente. Per esempio, quasi tutti i campi con valore 0 e 1 si possono considerare flag.

Figura 21-3
Impostazione del livello di misurazione per più campi

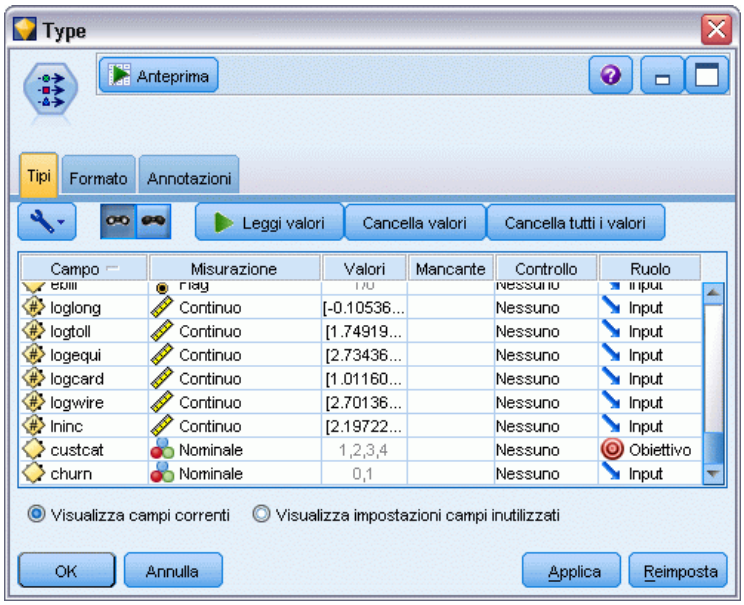


Suggerimento: se si desidera modificare le proprietà di più campi con valori simili (quali 0/1), fare clic sull'intestazione della colonna *Valori* per ordinare i campi in base al valore e tenere premuto il tasto Maiusc mentre si selezionano con il mouse o i tasti freccia tutti i campi da modificare. A questo punto, è possibile fare clic con il pulsante destro del mouse sulla selezione per modificare il livello di misurazione o altri attributi dei campi selezionati.

Si noti che *sesso* è considerato più correttamente un campo con un insieme di due valori, anziché un flag; pertanto, lasciare il suo valore Misurazione impostato su Nominale.

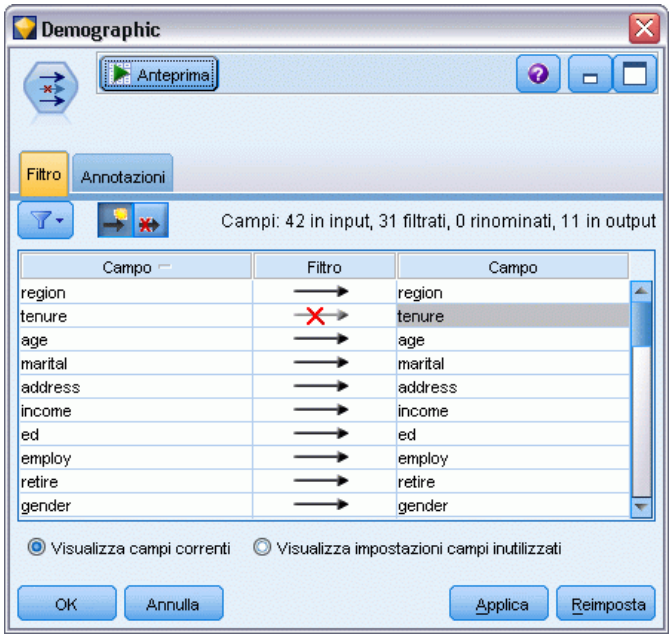
- Impostare il ruolo del campo *catcli* su Obiettivo. Tutti gli altri campi dovrebbero avere il ruolo impostato su Input.

Figura 21-4
Impostazione del ruolo del campo



Poiché questo esempio prende in esame dati demografici, utilizzare il nodo Filtro per includere solo i campi rilevanti (*regione*, *età*, *statciv*, *indirizzo*, *reddito*, *istruz*, *impiego*, *pensionato/a*, *sesso*, *residenza* e *catpers*). Ai fini di questa analisi, è possibile escludere gli altri campi.

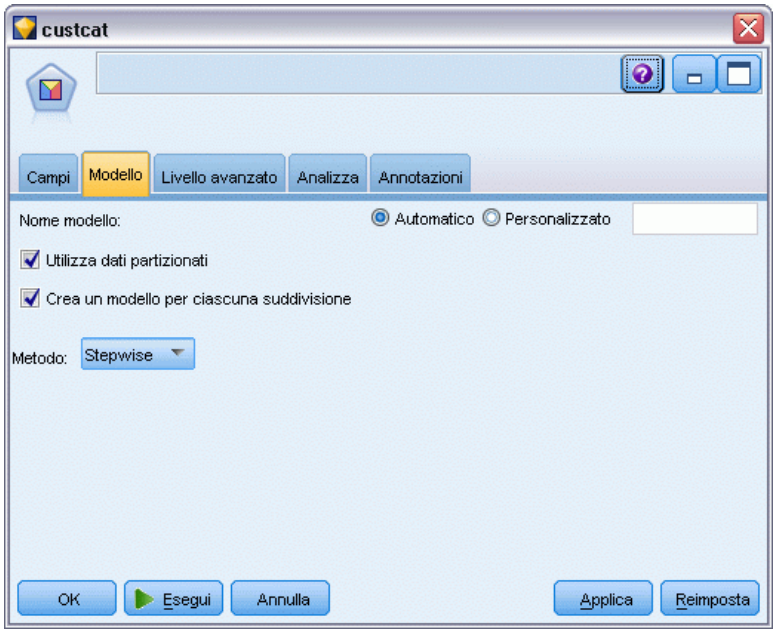
Figura 21-5
Filtro di campi demografici



In alternativa, anziché escludere tali campi, è possibile modificarne il ruolo in Nessuno oppure selezionare i campi da utilizzare nel nodo Modelli.

▶ Nel nodo Discriminante, fare clic sulla scheda Modello e selezionare il metodo Stepwise.

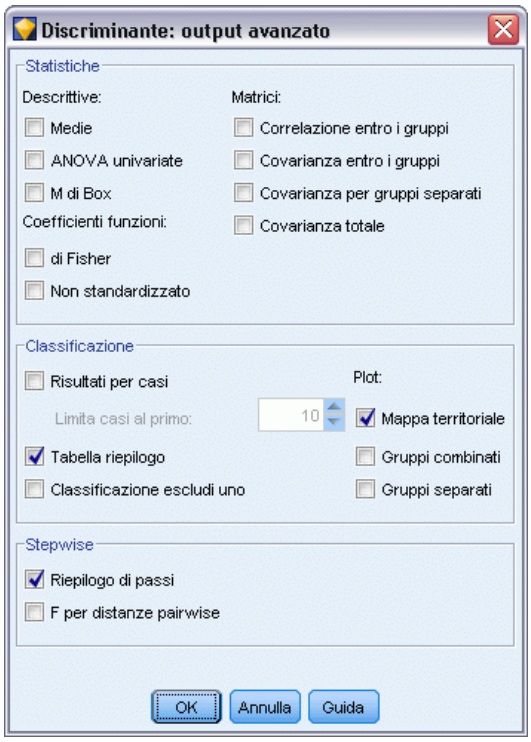
Figura 21-6
Scelta delle opzioni del modello



▶ Nella scheda Livello avanzato, impostare la modalità su Livello avanzato e fare clic su Output.

▶ Selezionare Tabella di riepilogo, Mappa territoriale e Riepilogo di passi nella finestra di dialogo Output avanzato, quindi fare clic su OK.

Figura 21-7
Scelta delle opzioni di output

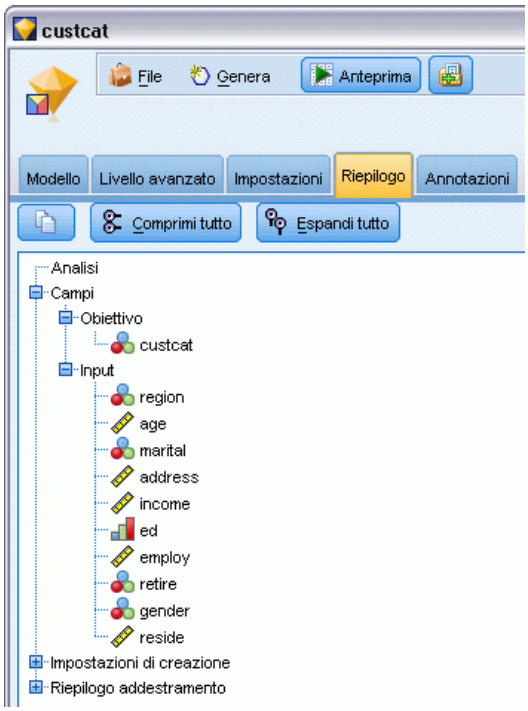


Esame del modello

▶ Fare clic su Esegui per creare il modello, che viene aggiunto allo stream e alla palette Modelli nell'angolo superiore destro. Per visualizzarne i dettagli, fare doppio clic sull'insieme di modelli nello stream.

La scheda Riepilogo, fra le altre cose, mostra il campo obiettivo e l'elenco completo dei campi di input (predittori) sottoposti all'analisi.

Figura 21-8
Riepilogo del modello che mostra i campi obiettivo e di input



Per informazioni sui risultati dell'analisi discriminante:

- Fare clic sulla scheda Avanzate.
- Fare clic sul pulsante "Avvia in un browser esterno" (subito sotto la scheda Modello) per visualizzare i risultati nel proprio browser Web.

Analisi discriminante stepwise

Figura 21-9
Variabili non in analisi, passaggio 0

Passo		Tolleranza	Tolleranza min.	F di inserimento	Lambda di Wilks
0	Age in years	1,000	1,000	7,521	,978
	Marital status	1,000	1,000	3,500	,990
	Years at current address	1,000	1,000	8,433	,975
	Household income in thousands	1,000	1,000	6,689	,980
	Level of education	1,000	1,000	61,454	,844
	Years with current employer	1,000	1,000	16,976	,951
	Retired	1,000	1,000	3,005	,991
	Gender	1,000	1,000	,373	,999
	Number of people in household	1,000	1,000	3,976	,988

Quando vi sono molti predittori, il metodo stepwise può essere utile per la selezione automatica delle variabili "ottimali" da usare nel modello. Il metodo stepwise inizia con un modello che non comprende alcun predittore. A ogni passaggio viene aggiunto al modello il predittore con il valore *F per inserimento* maggiore che supera il criterio di inserimento (per default, 3.84).

Figura 21-10
Variabili non in analisi, passaggio 3

Passo		Tolleranza	Tolleranza min.	F di inserimento	Lambda di Wilks
3	Age in years	,535	,535	,252	,795
	Marital status	,605	,593	1,507	,792
	Years at current address	,776	,771	3,514	,787
	Household income in thousands	,688	,657	,687	,794
	Retired	,917	,880	,353	,795
	Gender	,997	,931	,395	,795

Le variabili escluse dall'analisi nell'ultimo passaggio hanno tutte valori *F per inserimento* minori di 3.84, perciò non ne vengono aggiunte altre.

Figura 21-11
Variabili nell'analisi

Passo		Tolleranza	F di rimozione	Lambda di Wilks
1	Level of education	1,000	61,454	
2	Level of education	,953	59,108	,951
	Years with current employer	,953	14,933	,844
3	Level of education	,951	60,046	,940
	Years with current employer	,934	15,824	,834
	Number of people in household	,979	4,841	,807

Questa tabella mostra le statistiche relative alle variabili presenti nell'analisi a ogni passaggio. La *Tolleranza* è la proporzione della varianza di una variabile non spiegata dalle altre variabili indipendenti nell'equazione. Una variabile con una tolleranza molto bassa apporta poche informazioni a un modello e può provocare problemi computazionali.

I valori *F per rimozione* sono utili per descrivere cosa avviene se una variabile viene rimossa dal modello corrente (dato che le altre variabili rimangono). Il valore *F per rimozione* per l'inserimento della variabile è uguale al valore *F per inserimento* nel passaggio precedente (riportato nella tabella Variabili non in analisi).

Nota precauzionale relativa ai metodi stepwise

I metodi stepwise sono utili ma presentano alcuni limiti. Poiché i metodi stepwise selezionano modelli basati unicamente sul merito statistico, potrebbero scegliere dei predittori privi di **significatività pratica**. Se si ha una certa esperienza con i dati e si nutrono aspettative su quali predittori siano importanti, è opportuno utilizzare queste conoscenze ed evitare i metodi stepwise. Se invece si hanno molti predittori e non si sa da dove iniziare, è meglio eseguire un'analisi stepwise e regolare il modello selezionato anziché partire senza alcun modello.

Verifica dell'adattamento del modello

Figura 21-12
Autovalori

Funzione	Autovalore	% di varianza	% cumulata	Correlazione canonica
1	,198(a)	80,2	80,2	,407
2	,048(a)	19,4	99,6	,214
3	,001(a)	,4	100,0	,031

a. Per l'analisi sono state usate le prime 3 funzioni discriminanti canoniche.

Quasi tutta la varianza spiegata dal modello è dovuta alle prime due funzioni discriminanti. Vengono adattate automaticamente tre funzioni ma, dato il suo autovalore minimo, è possibile ignorare tranquillamente la terza.

Figura 21-13
Lambda di Wilks

Test di funzioni	Lambda di Wilks	Chi-quadrato	df	Sig.
Da 1 a 3	,796	227,345	9	,000
Da 2 a 3	,953	47,486	4	,000
3	,999	,929	1	,335

Il lambda di Wilks conferma che solo le prime due funzioni sono utili. Per ogni insieme di funzioni, questo test verifica l'ipotesi che le medie delle funzioni elencate siano uguali in tutti i gruppi. Il test della funzione 3 ha un valore di significatività maggiore di 0.10, pertanto questa funzione dà uno scarso contributo al modello.

Matrice della struttura

Figura 21-14
Matrice della struttura

	Funzione		
	1	**2**	**3**
Level of education	,966(*)	-,090	-,244
Geographic indicator(a)	-,073(*)	,002	-,049
Years with current employer	-,182	,964(*)	-,193
Age in years(a)	-,162	,598(*)	-,285
Household income in thousands(a)	,109	,514(*)	-,190
Years at current address(a)	-,151	,394(*)	-,214
Retired(a)	-,108	,230(*)	-,137
Gender(a)	,008	,054(*)	,009
Number of people in household	,232	,097	,968(*)
Marital status(a)	,132	,134	,600(*)
Correlazioni comuni entro gruppi tra variabili discriminanti e funzioni discriminanti canoniche standardizzate Variabili ordinate in base alla dimensione assoluta della correlazione entro la funzione.			
*. Correlazione assoluta più grande tra ciascuna variabile e qualsiasi funzione discriminante			
a. Questa variabile non viene usata nell'analisi.			

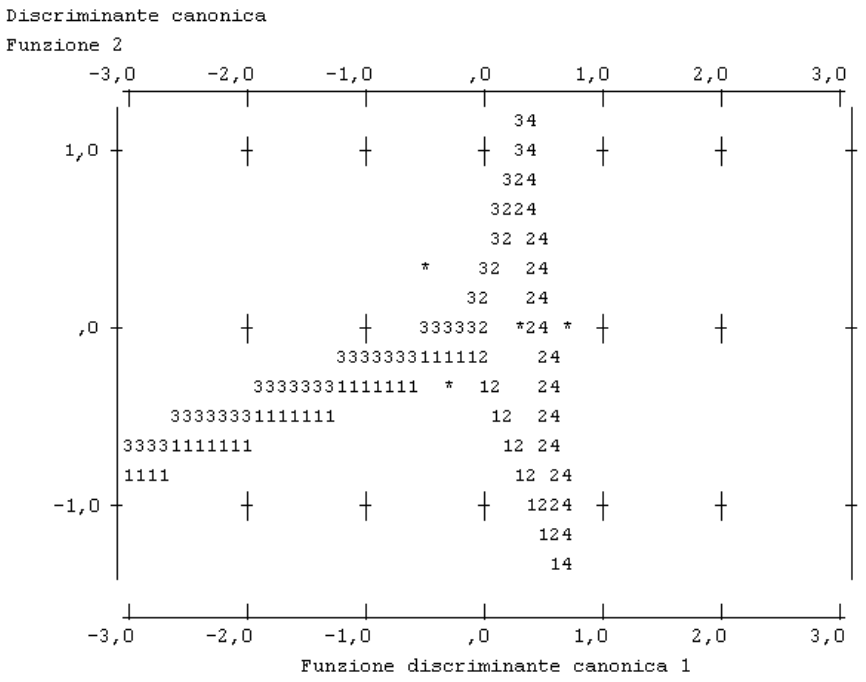
Quando è presente più di una funzione discriminante, un segno di asterisco (*) contrassegna la correlazione assoluta massima di ogni variabile con una delle funzioni canoniche. All'interno di ogni funzione, queste variabili contrassegnate vengono quindi ordinate in base alle dimensioni della correlazione.

- Il *Grado di istruzione* è correlato più strettamente alla prima funzione ed è l'unica variabile più strettamente correlata a questa funzione.

- *Anni di permanenza nell'impiego attuale*, *Età in anni*, *Reddito familiare in migliaia*, *Anni di residenza all'indirizzo corrente*, *Pensionamento* e *Sesso* sono più strettamente correlati alla seconda funzione, sebbene *Sesso* e *Pensionamento* siano correlati più debolmente rispetto agli altri. Le altre variabili contrassegnano questa funzione come una funzione di "stabilità".
- *Membri del nucleo* e *Stato civile* sono più strettamente correlati alla terza funzione discriminante, ma si tratta di una funzione priva di utilità, pertanto anche questi predittori sono praticamente privi di utilità.

Mappa territoriale

Figura 21-15
Mappa territoriale



La mappa territoriale aiuta a studiare le relazioni tra i gruppi e le funzioni discriminanti. Combinata con i risultati della matrice della struttura, fornisce un'interpretazione grafica della relazione tra predittori e gruppi. La prima funzione, mostrata sull'asse orizzontale, separa il gruppo 4 (clienti *Servizio totale*) dagli altri. Poiché il *Grado di istruzione* ha una forte correlazione positiva con la prima funzione, i clienti *Servizio totale* sono, in generale, quelli con il livello di istruzione più elevato. La seconda funzione separa i gruppi 1 e 3 (clienti *Servizio base* e *Servizio avanzato*). I clienti *Servizio avanzato* tendono ad aver lavorato più a lungo e sono più anziani dei clienti *Servizio base*. I clienti *E-service* non sono separati nettamente dagli altri, sebbene la mappa suggerisca che tendono ad avere un buon livello di istruzione con una moderata esperienza lavorativa.

In generale, la vicinanza dei baricentri dei gruppi, contrassegnati dagli asterischi (*), alle linee territoriali suggerisce che la separazione tra i gruppi non è molto netta.

Solo le prime due funzioni discriminanti sono rappresentate graficamente, ma poiché si è scoperto che la terza funzione è scarsamente significativa, la mappa territoriale offre una visione completa del modello discriminante.

Risultati di classificazione

Figura 21-16
Risultati di classificazione

		Customer category	Gruppo di appartenenza previsto				Totali
			Basic service	E-service	Plus service	Total service	
Originale	Conteggio	Basic service	125	11	61	69	266
		E-service	49	15	58	95	217
		Plus service	102	14	112	53	281
		Total service	40	16	37	143	236
	%	Basic service	47,0	4,1	22,9	25,9	100,0
		E-service	22,6	6,9	26,7	43,8	100,0
		Plus service	36,3	5,0	39,9	18,9	100,0
		Total service	16,9	6,8	15,7	60,6	100,0
a. 39,5% di casi raggruppati originali classificati correttamente.							

Dal lambda di Wilks, si sa che il modello consente di ottenere più di semplici ipotesi, ma è necessario passare ai risultati di classificazione per determinarlo con maggior precisione. Dati i dati osservati, il modello "nullo" (ovvero, il modello senza predittori) classificherebbe tutti i clienti nel gruppo modale *Servizio avanzato*. Pertanto, il modello nullo sarebbe corretto 281/1000 = 28.1% delle volte. Il modello ottiene l'11.4% in più o il 39.5% dei clienti. In particolare, il modello eccelle nell'identificazione dei clienti *Servizio totale*. Tuttavia, la sua performance è incredibilmente scarsa nella classificazione dei clienti *E-service*. Per separare questi clienti potrebbe essere necessario individuare un altro predittore.

Riepilogo

È stato creato un modello discriminante che classifica i clienti in uno dei quattro gruppi predefiniti di "utilizzo del servizio", basati sui dati demografici di ogni cliente. Con la matrice della struttura e la mappa territoriale sono state identificate le variabili più utili per la segmentazione della base di clienti. Infine, i risultati di classificazione mostrano che il modello è scarsamente efficace per la classificazione dei clienti *E-service*. È necessario effettuare ulteriori ricerche per determinare un'altra variabile predittore in grado di classificare meglio questi clienti ma, a seconda di ciò che si desidera prevedere, il modello potrebbe essere perfettamente adeguato alle necessità. Per esempio, se non si è interessati all'individuazione dei clienti *E-service*, il modello potrebbe essere sufficientemente preciso per le proprie esigenze. Ciò potrebbe avvenire, per esempio, se E-service fosse un prodotto civetta che genera scarsi profitti. Se, per esempio, il ROI massimo provenisse dai clienti *Servizio avanzato* o *Servizio totale*, il modello potrebbe fornire le informazioni necessarie.

Si noti inoltre che questi risultati si basano soltanto sui dati di addestramento. Per un'indicazione dell'efficacia con cui il modello potrà essere esteso ad altri dati, utilizzare un nodo Partizione per estrarre un sottoinsieme di record da utilizzare per test e validazione. Per ulteriori informazioni, vedere l'argomento Nodo Partizione in il capitolo 4 in *IBM SPSS Modeler 15 Nodi di input, elaborazione e output*.

Per una spiegazione dei fondamenti matematici dei metodi di modellazione utilizzati in IBM® SPSS® Modeler vedere il manuale SPSS Modeler Algorithms Guide, disponibile nella directory *\Documentation* del disco di installazione.

Capitolo
22

Analisi dei dati di sopravvivenza censurati per intervalli (modelli lineari generalizzati)

Quando si analizzano i dati di sopravvivenza con la censura per intervalli, ovvero quando il tempo esatto di un evento di interesse non è noto ma si sa solo che è avvenuto entro un determinato intervallo, l'applicazione del modello di Cox ai rischi degli eventi negli intervalli consente di ottenere un modello di regressione doppia logaritmica complementare.

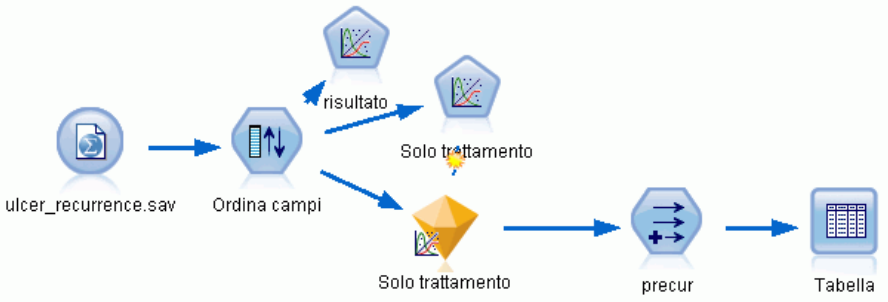
Le informazioni parziali provenienti da uno studio progettato per confrontare l'efficacia di due terapie per la prevenzione delle recidive nei malati di ulcera sono raccolte in *ulcer_recurrence.sav*. Questo insieme di dati è stato presentato e analizzato in un'altra sezione. Utilizzando i modelli lineari generalizzati, è possibile replicare i risultati dei modelli di regressione doppia logaritmica complementare.

In questo esempio viene utilizzato lo stream denominato *ulcer_genlin.str*, che fa riferimento al file di dati denominato *ulcer_recurrence.sav*. Il file di dati si trova nella cartella *Demos* e il file di stream nella sottocartella *streams*. Per ulteriori informazioni, vedere l'argomento Cartella Demos in il capitolo 1 in *Manuale dell'utente di IBM SPSS Modeler 15*.

Creazione del flusso

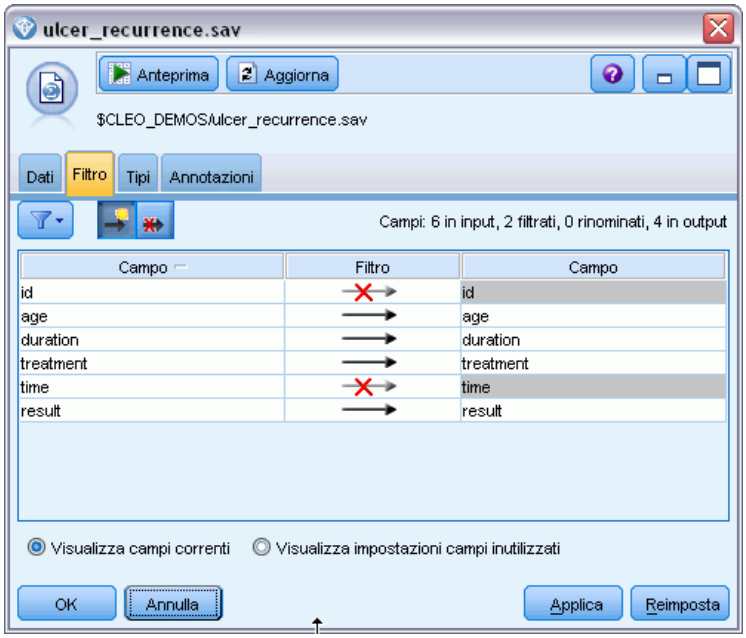
- Aggiungere un nodo di input File Statistics che punta a *ulcer_recurrence.sav* nella cartella *Demos*.

Figura 22-1
Esempio di stream per prevedere le recidive nei malati di ulcera



- Nella scheda Filtro del nodo di input, filtrare *id* e *time*.

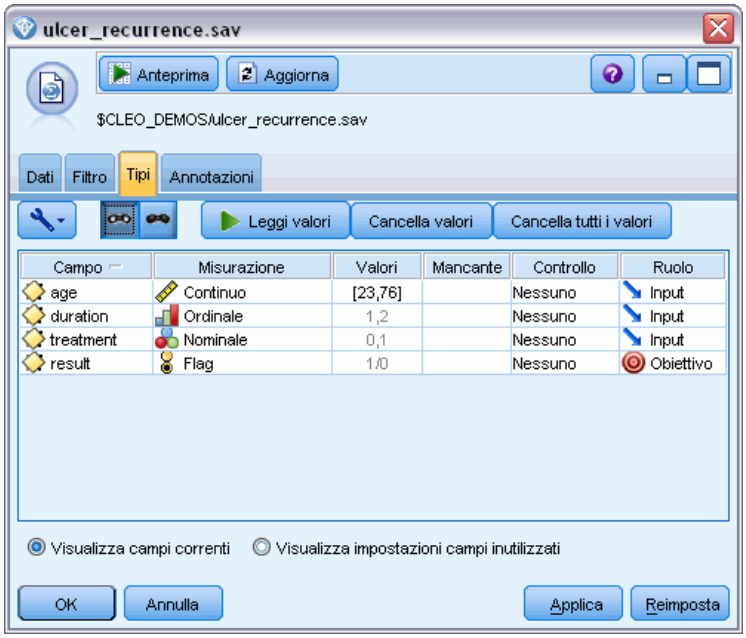
Figura 22-2
Filtro dei campi non desiderati



- Nella scheda Tipi del nodo di input, impostare il ruolo del campo *result* su Obiettivo e impostarne il livello di misurazione su Flag. Un risultato di 1 indica che l'ulcera si è ripresentata. Tutti gli altri campi dovrebbero avere il ruolo impostato su Input.

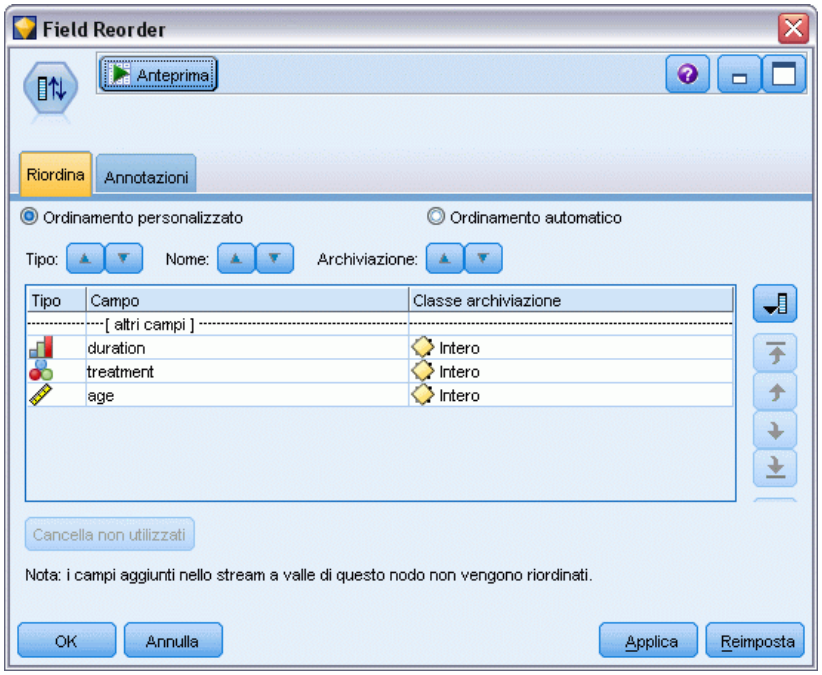
▶ Fare clic su Read Values (Leggi i valori) per creare un'istanza dei dati.

Figura 22-3
Impostazione del ruolo del campo



- Aggiungere un nodo Riordina campi e specificare *duration*, *treatment* e *age* come ordine di input. Ciò determina l'ordine nel quale i campi vengono inseriti nel modello e aiuta l'utente a replicare i risultati di Collett.

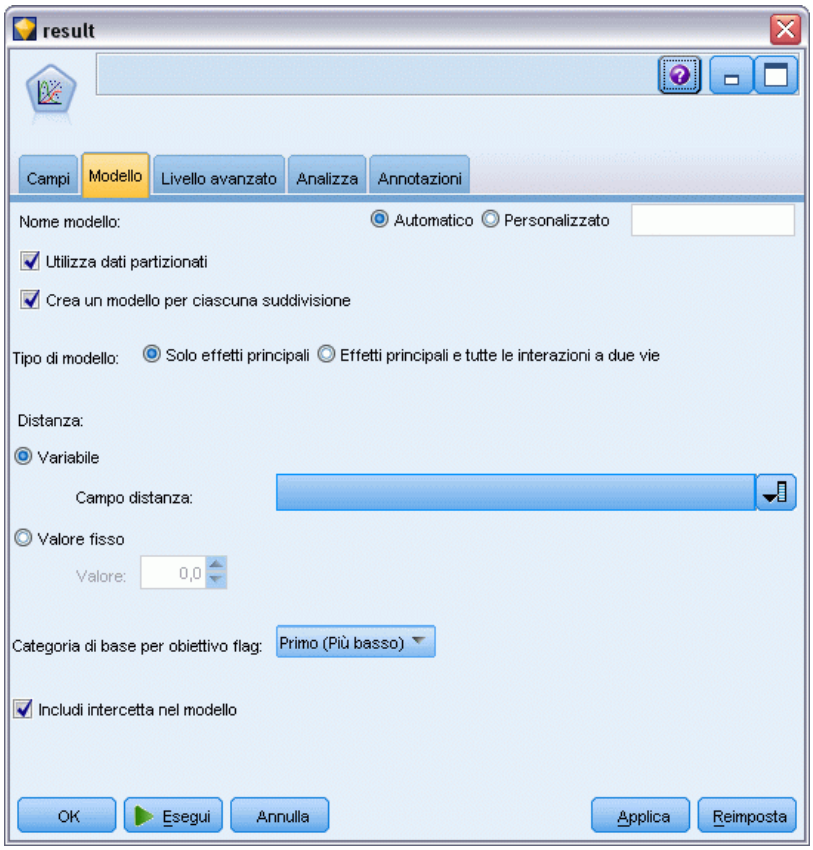
Figura 22-4
Riordinamento dei campi per far sì che vengano inseriti nel modello nel modo desiderato



- Collegare un nodo GenLin al nodo di input; nel nodo GenLin, fare clic sulla scheda Modello.
- Selezionare Prima (più bassa) come categoria di riferimento per l'obiettivo. Ciò indica che la seconda categoria è l'evento di interesse e che i suoi effetti sul modello riguardano l'interpretazione delle stime dei parametri. Un predittore continuo con un coefficiente positivo indica una maggiore probabilità di recidiva con valori crescenti del predittore; categorie di un

predittore nominale con coefficienti più alti indicano maggiore probabilità di recidiva rispetto ad altre categorie dell'insieme.

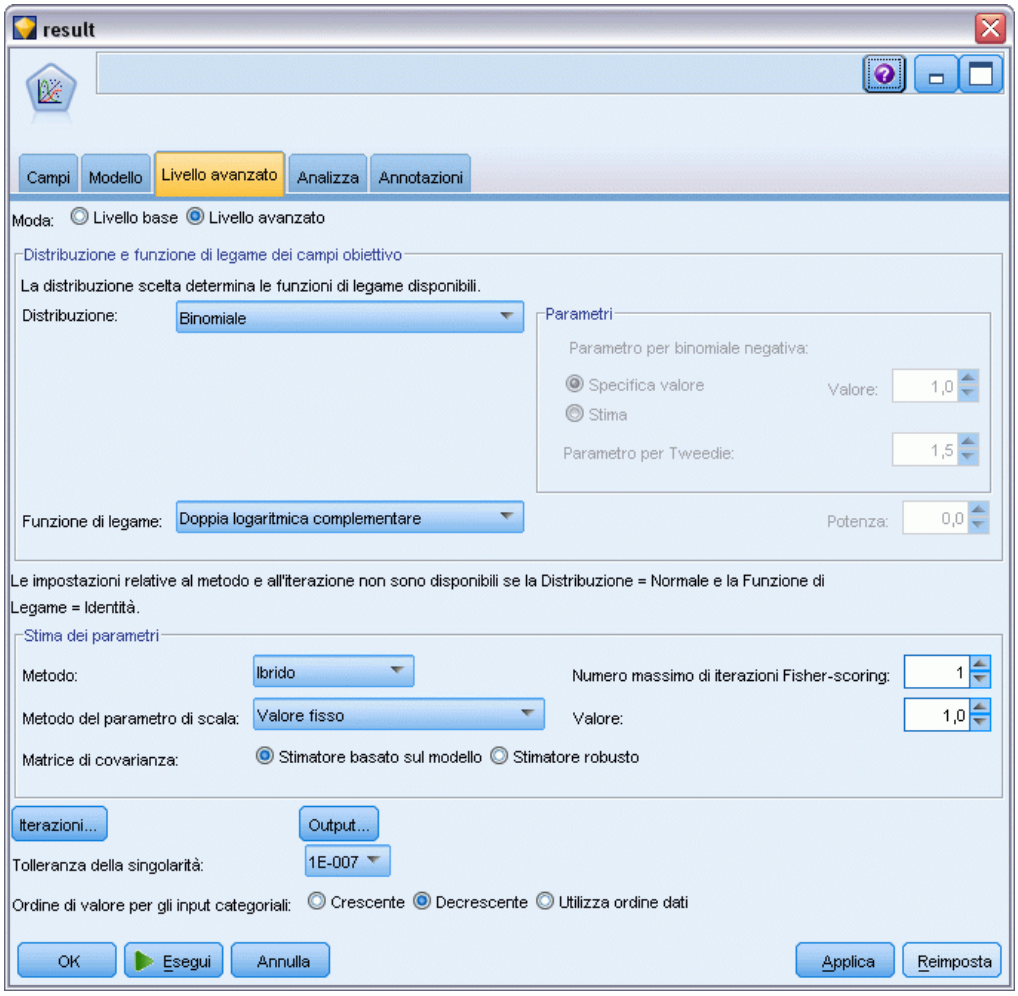
Figura 22-5
Scelta delle opzioni del modello



- Fare clic sulla scheda Livello avanzato e selezionare Livello avanzato per attivare le opzioni di modellazione avanzata.
- Selezionare Binomiale come distribuzione e Doppia logaritmica complementare come funzione di legame.
- Selezionare Valore fisso come metodo per la stima del parametro di scala e lasciare il valore di default impostato su 1.0.

- Selezionare Decrescente come ordine delle categorie per i fattori. Questo indica che la prima categoria di ogni fattore sarà la relativa categoria di riferimento; l'effetto di questa selezione sul modello è nell'interpretazione delle stime dei parametri.

Figura 22-6
Scelta delle opzioni Expert



- Eseguire lo stream per creare l'insieme di modelli, che viene aggiunto allo stream e alla palette Modelli nell'angolo superiore destro. Per visualizzare i dettagli dei modelli, fare clic con il pulsante destro del mouse sull'insieme di modelli e selezionare Modifica o Sfoglia.

Test degli effetti del modello

Figura 22-7
Test degli effetti del modello per il modello effetti principali

Sorgente	Tipo III		
	Chi-quadrato di Wald	df	Sig.
(Intercetta)	,536	1	,464
duration	,003	1	,958
treatment	,382	1	,537
age	,358	1	,550
Variabile dipendente: ResultModello: (Intercetta), duration, treatment, age			

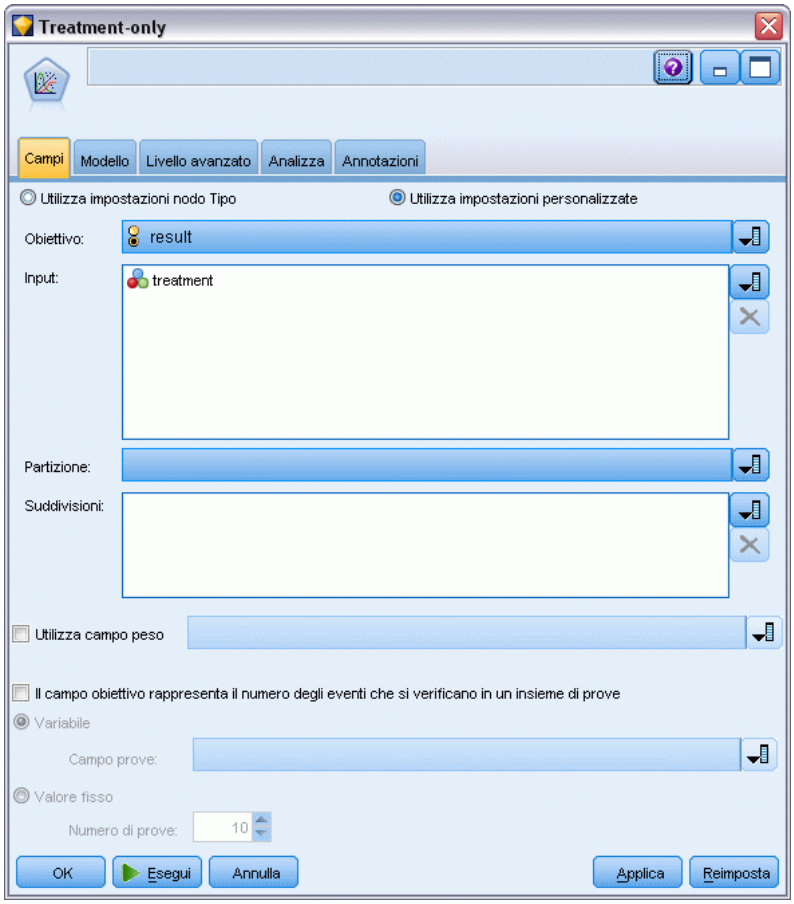
Nessuno degli effetti del modello è statisticamente significativo; tuttavia, qualsiasi differenza osservabile negli effetti del trattamento è di interesse clinico, per cui si adatterà un modello ridotto con il solo trattamento come termine del modello.

Adattamento del modello solo trattamento

- Nella scheda Campi del nodo GenLin, fare clic su Utilizza impostazioni personalizzate.
- Selezionare *result* come obiettivo.

▶ Selezionare *treatment* come unico input.

Figura 22-8
Scelta delle opzioni dei campi



▶ Eseguire lo stream e aprire l'insieme di modelli generato.

Nell'insieme di modelli selezionare la scheda Opzioni avanzate e scorrerla fino in fondo.

Stime di parametri

Figura 22-9
Stime parametri per il modello solo trattamento

Parametro	B	Deviazione standard Errore	95% Intervallo di confidenza di Wald		Test dell'ipotesi		
			Inferiore	Superiore	Chi-quadrato di Wald	df	Sig.
(Intercetta)	-1,442	,5012	-2,425	-,460	8,282	1	,004
[treatment=1]	,378	,6288	-,855	1,610	,361	1	,548
[treatment=0]	0(a)	.	.	.	.	.	.
(Scala)	1(b)						

Variabile dipendente: ResultModello: (Intercetta), treatment, offset = 0

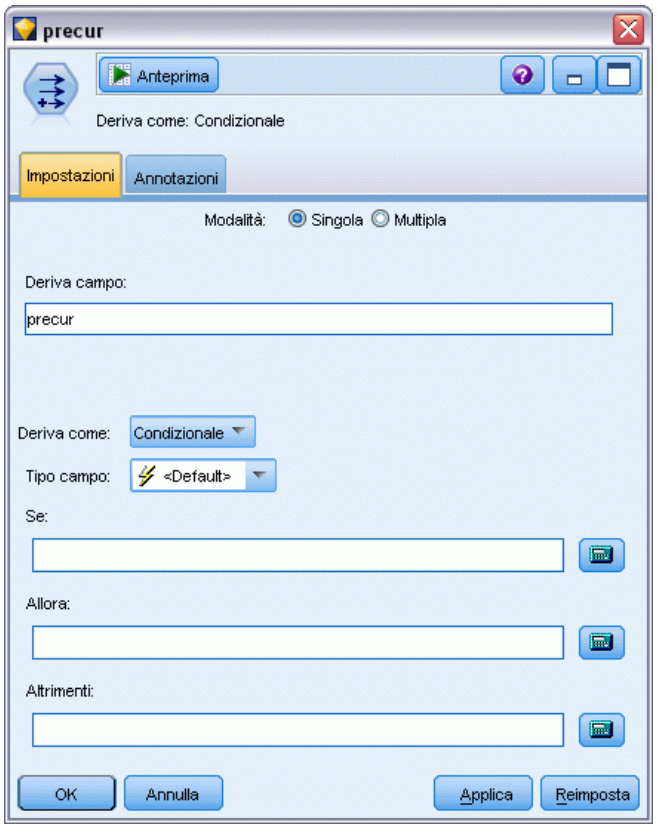
a. Impostato su zero poiché questo parametro è ridondante.

b. Fissato al valore visualizzato.

L'effetto del trattamento (la differenza del predittore lineare tra i due livelli di trattamento, ovvero, il coefficiente di *[treatment=1]*) non è ancora statisticamente significativo, ma suggerisce solo che il trattamento *A [treatment=0]* potrebbe essere migliore di *B [treatment=1]* perché la stima dei parametri per il trattamento *B* è maggiore che per *A* ed è pertanto associata a una maggiore probabilità di recidiva nei primi 12 mesi. Il predittore lineare (intercetta + effetto del trattamento) è una stima di $\log(-\log(1-P(recur_{12,t}))$, dove $P(recur_{12,t})$ è la probabilità di recidiva a 12 mesi per il trattamento t(=*A* o *B*). Queste probabilità previste vengono generate per ogni osservazione nell'insieme di dati.

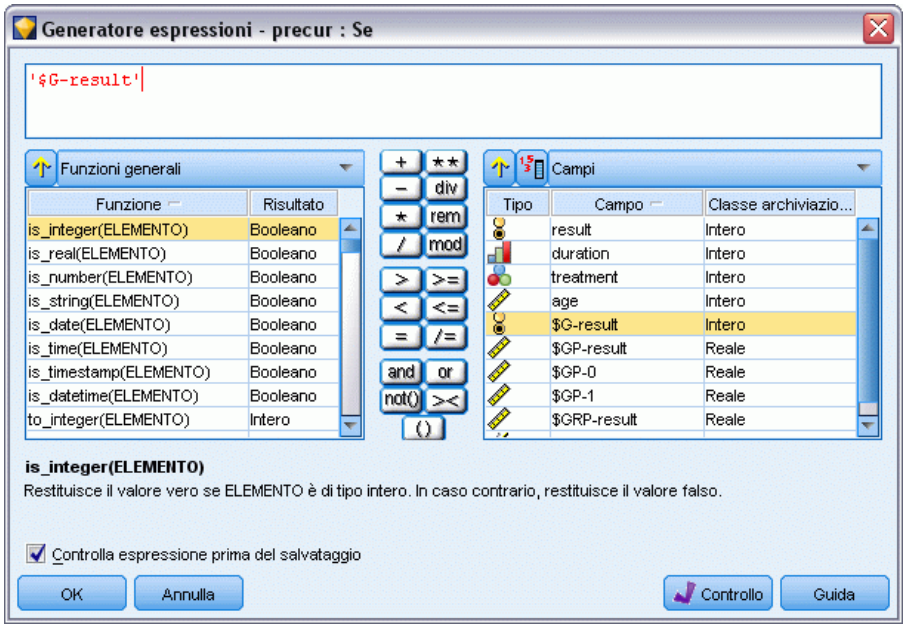
Recidive previste e probabilità di sopravvivenza

Figura 22-10
Opzioni della scheda Impostazioni per il nodo Nuovo campo



- Per ogni paziente, il modello calcola il punteggio del risultato previsto e la probabilità di tale risultato previsto. Per visualizzare le probabilità di recidiva previste, copiare il modello generato nella palette e collegare un nodo Nuovo campo.
- Nella scheda Impostazioni, immettere precur come nuovo campo.
- Scegliere la derivazione Condizionale.
- Fare clic sul pulsante del calcolatore per aprire il Generatore di espressioni per la condizione Se.

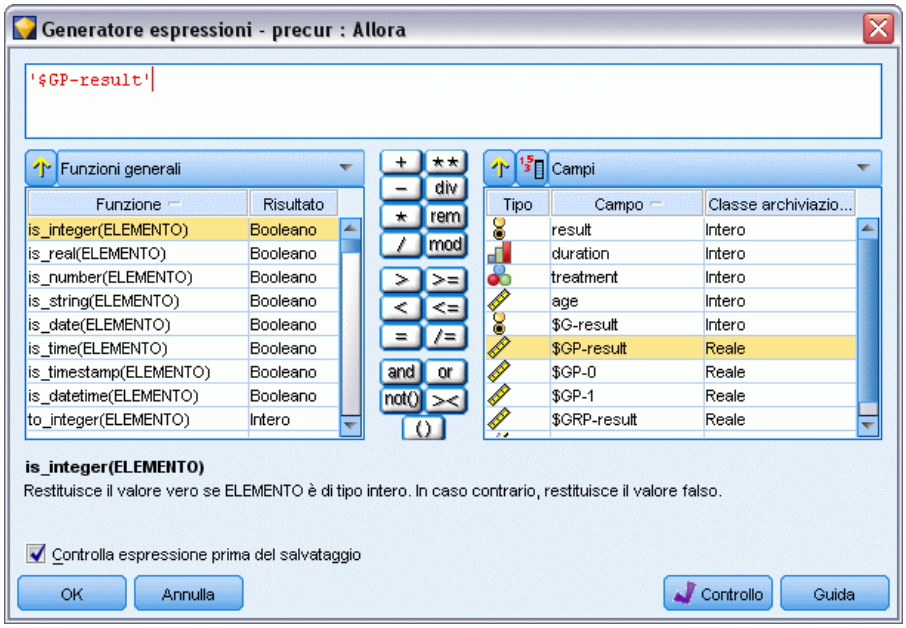
Figura 22-11
Nodo Nuovo campo: Generatore di espressioni per la condizione Se



- Inserire il campo *$G-risultato* nell'espressione.
- Fare clic su OK.

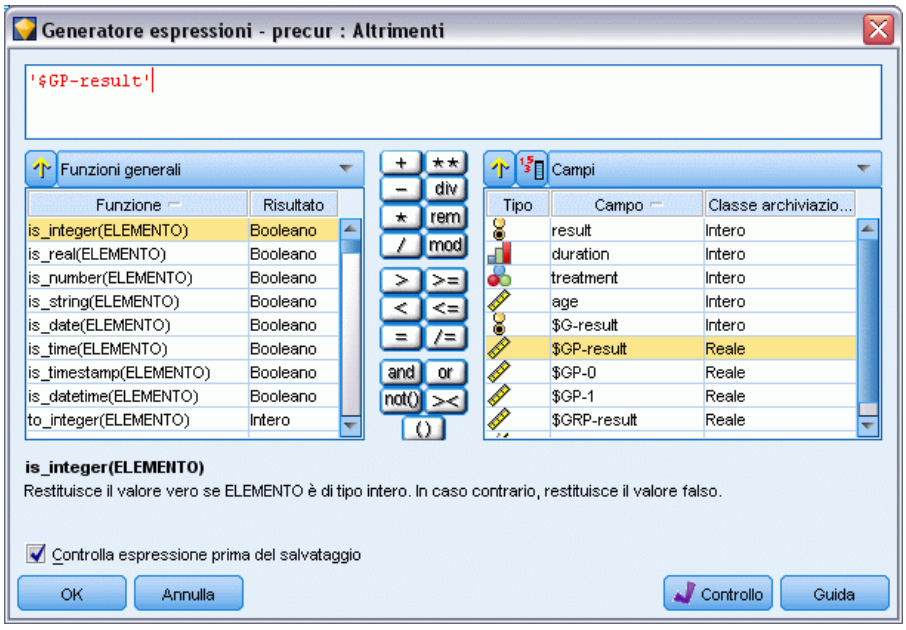
Il nuovo campo *precur* prende solo il valore dell'espressione Allora quando *$G-risultato* è uguale a 1 e il valore dell'espressione Altrimenti quando è 0.

Figura 22-12
Nodo Nuovo campo: Generatore di espressioni per l'espressione Allora



- Fare clic sul pulsante del calcolatore per aprire il Generatore di espressioni per l'espressione Allora.
- Inserire il campo *$GP-risultato* nell'espressione.
- Fare clic su OK.

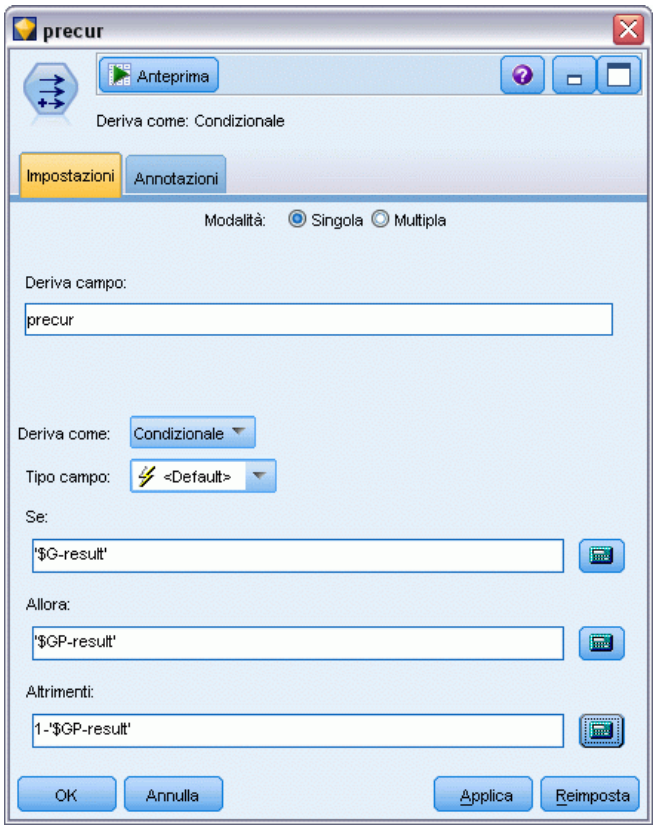
Figura 22-13
Nodo Nuovo campo: Generatore di espressioni per l'espressione Else



- Fare clic sul pulsante del calcolatore per aprire il Generatore di espressioni per l'espressione Else.

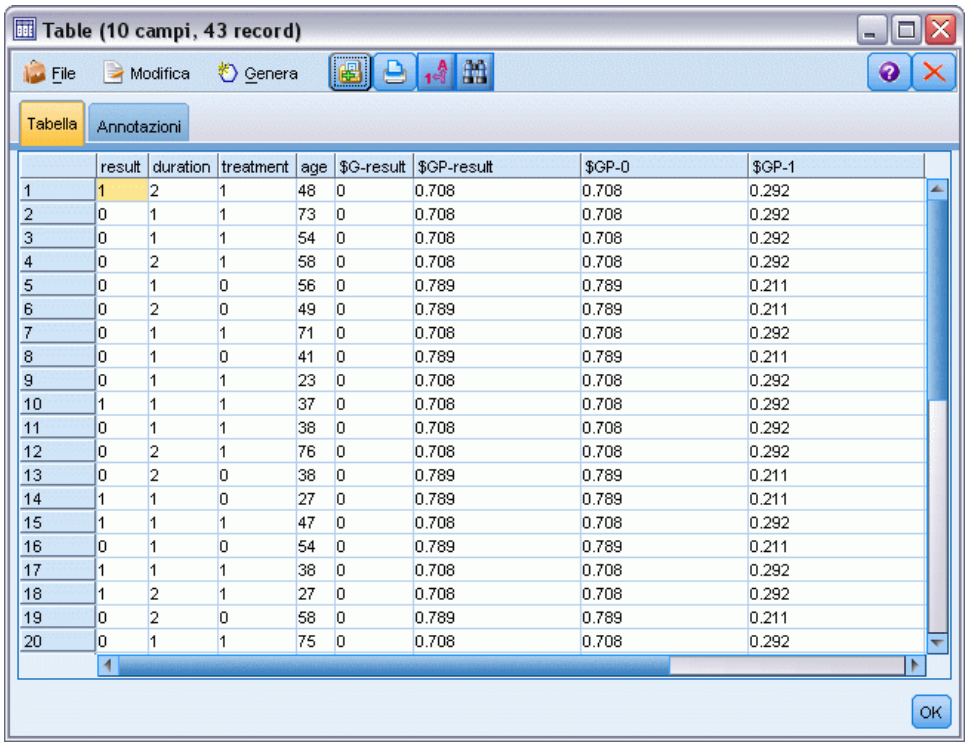
- Immettere 1- nell'espressione e quindi inserire il campo *$GP-risultato* nell'espressione.
- Fare clic su OK.

Figura 22-14
Opzioni della scheda Impostazioni per il nodo Nuovo campo



- Collegare un nodo Tabella al nodo Nuovo campo e avviarne l'esecuzione.

Figura 22-15
Probabilità attese



	result	duration	treatment	age	$G-result	$GP-result	$GP-0	$GP-1
1	1	2	1	48	0	0.708	0.708	0.292
2	0	1	1	73	0	0.708	0.708	0.292
3	0	1	1	54	0	0.708	0.708	0.292
4	0	2	1	58	0	0.708	0.708	0.292
5	0	1	0	56	0	0.789	0.789	0.211
6	0	2	0	49	0	0.789	0.789	0.211
7	0	1	1	71	0	0.708	0.708	0.292
8	0	1	0	41	0	0.789	0.789	0.211
9	0	1	1	23	0	0.708	0.708	0.292
10	1	1	1	37	0	0.708	0.708	0.292
11	0	1	1	38	0	0.708	0.708	0.292
12	0	2	1	76	0	0.708	0.708	0.292
13	0	2	0	38	0	0.789	0.789	0.211
14	1	1	0	27	0	0.789	0.789	0.211
15	1	1	1	47	0	0.708	0.708	0.292
16	0	1	0	54	0	0.789	0.789	0.211
17	1	1	1	38	0	0.708	0.708	0.292
18	1	2	1	27	0	0.708	0.708	0.292
19	0	2	0	58	0	0.789	0.789	0.211
20	0	1	1	75	0	0.708	0.708	0.292

Vi è una probabilità prevista di 0.211 che i pazienti assegnati al trattamento *A* presentino una recidiva nei primi 12 mesi; 0.292 per il trattamento *B*. Si noti che $1-P(recur_{12,\,t})$ è la probabilità di sopravvivenza a 12 mesi, che può essere di maggior interesse per le analisi di sopravvivenza.

Modellazione della probabilità di recidiva per periodo

Un problema del modello nella sua forma attuale è il fatto che non tiene conto delle informazioni raccolte al primo esame; ovvero, del fatto che molti pazienti non hanno presentato recidive nei primi sei mesi. Un modello "migliore" comporterebbe la modellazione di una risposta binaria che registri il verificarsi o meno dell'evento durante ogni intervallo. L'adattamento di questo modello richiede una ricostruzione dell'insieme di dati originale, reperibile in *ulcer_recurrence_recoded.sav*. Per ulteriori informazioni, vedere l'argomento Cartella Demos in il capitolo 1 in *Manuale dell'utente di IBM SPSS Modeler 15*. Questo file contiene due variabili aggiuntive:

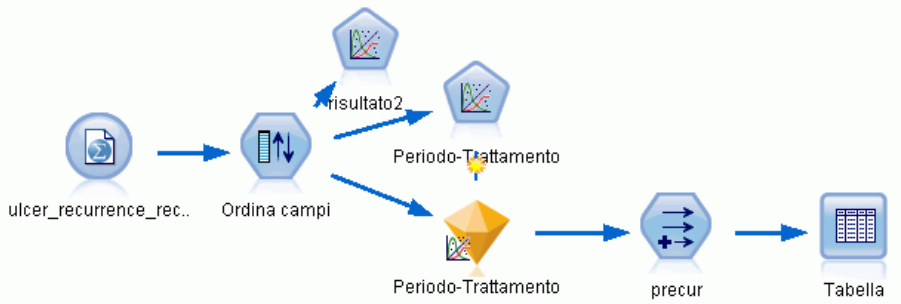
- *Periodo*, che registra se il caso corrisponde al primo o al secondo periodo di esame.
- *Risultato per periodo*, che registra se vi è stata recidiva per un determinato paziente nel periodo dato.

Ogni caso originale (paziente) contribuisce con un caso per ogni intervallo nel quale rimane nell'insieme di rischio. Così, per esempio, il paziente 1 contribuisce con due casi; uno per il primo periodo di esame in cui non vi è stata recidiva e uno per il secondo periodo di esame in cui è stata

riscontrata una recidiva. Il paziente 10, invece, contribuisce con un unico caso perché è stata riscontrata una recidiva nel primo periodo. I pazienti 16, 28 e 34 sono stati esclusi dallo studio dopo sei mesi e contribuiscono pertanto con un unico caso al nuovo insieme di dati.

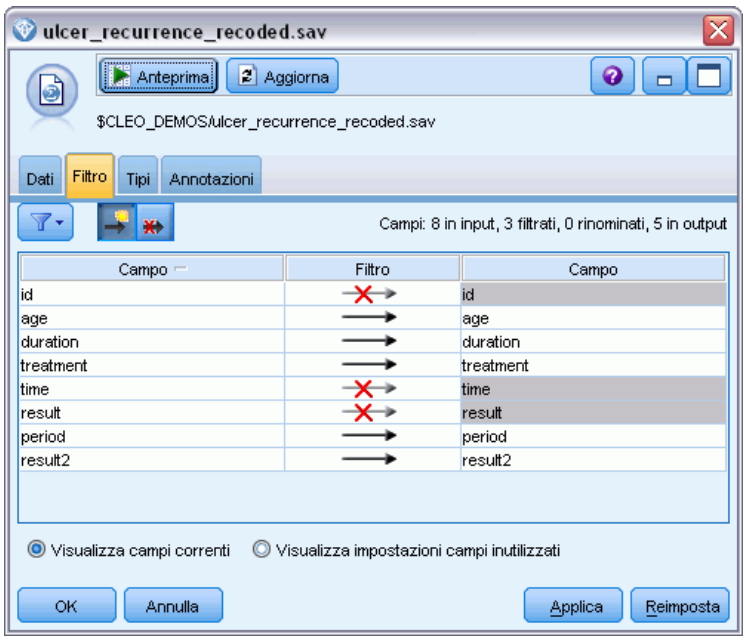
- Aggiungere un nodo di input File Statistics che punta a *ulcer_recurrence_recoded.sav* nella cartella *Demos*.

Figura 22-16
Esempio di stream per prevedere le recidive nei malati di ulcera



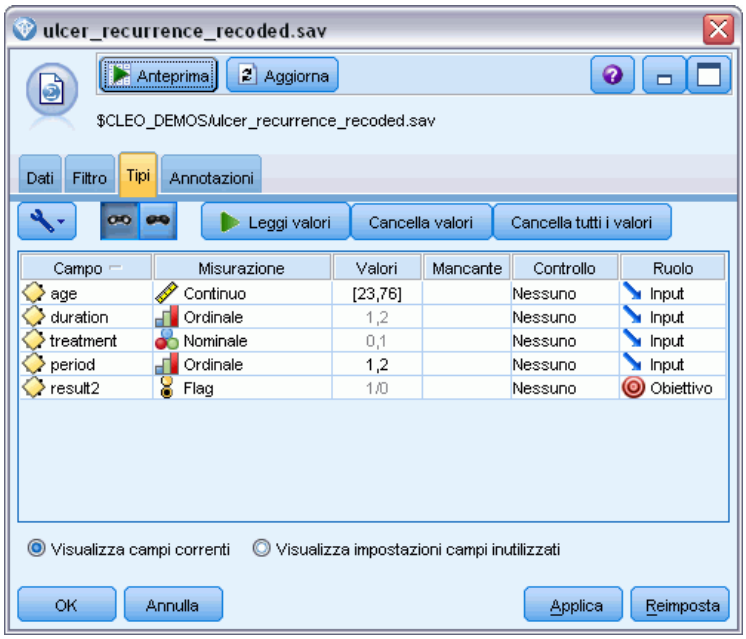
- Nella scheda Filtro del nodo di input, filtrare *id*, *time* e *result*.

Figura 22-17
Filtro dei campi non desiderati



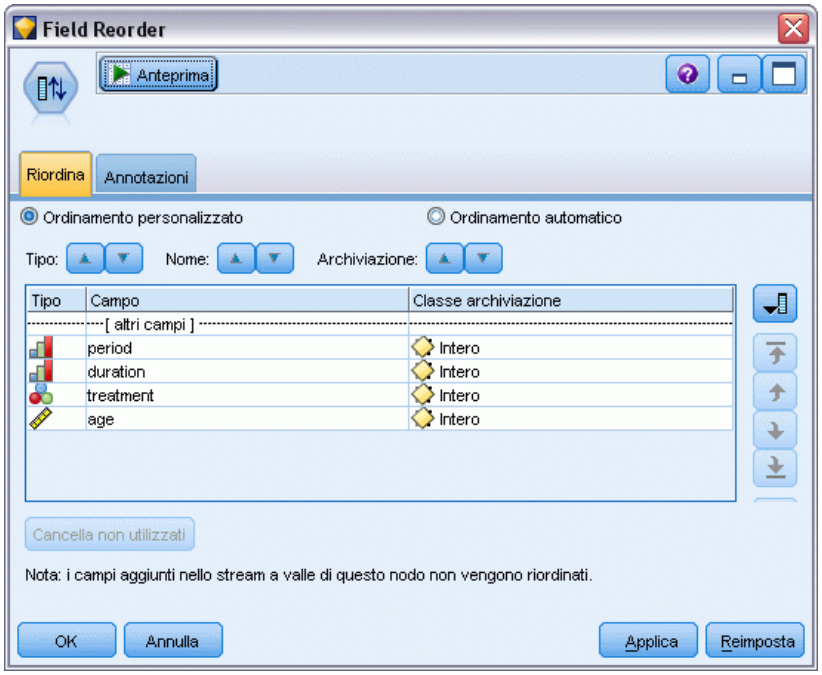
- Nella scheda Tipi del nodo di input, impostare il ruolo del campo *result2* su Obiettivo e impostarne il livello di misurazione su Flag. Tutti gli altri campi dovrebbero avere il ruolo impostato su Input.

Figura 22-18
Impostazione del ruolo del campo



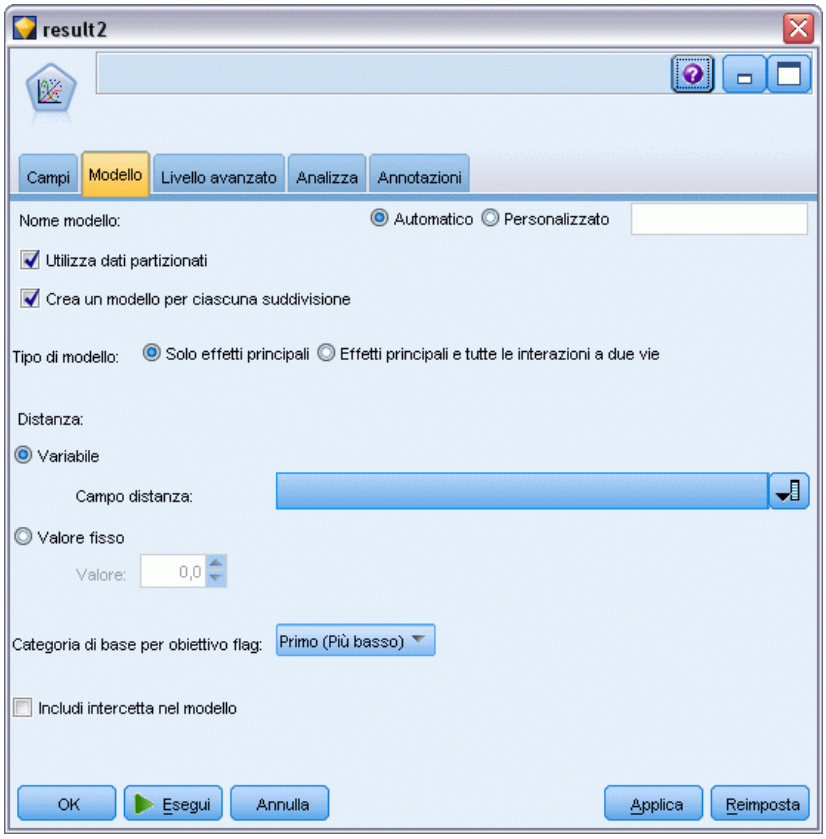
- Aggiungere un nodo Riordina campi e specificare *period*, *duration*, *treatment* e *age* come ordine di input. Rendendo *period* il primo input (e non includendo il termine intercetta nel modello) è possibile adattare un insieme completo di variabili fittizie per rilevare gli effetti del periodo.

Figura 22-19
Riordinamento dei campi per far sì che vengano inseriti nel modello nel modo desiderato



▶ Nel nodo GenLin, fare clic sulla scheda Modello.

Figura 22-20
Scelta delle opzioni del modello

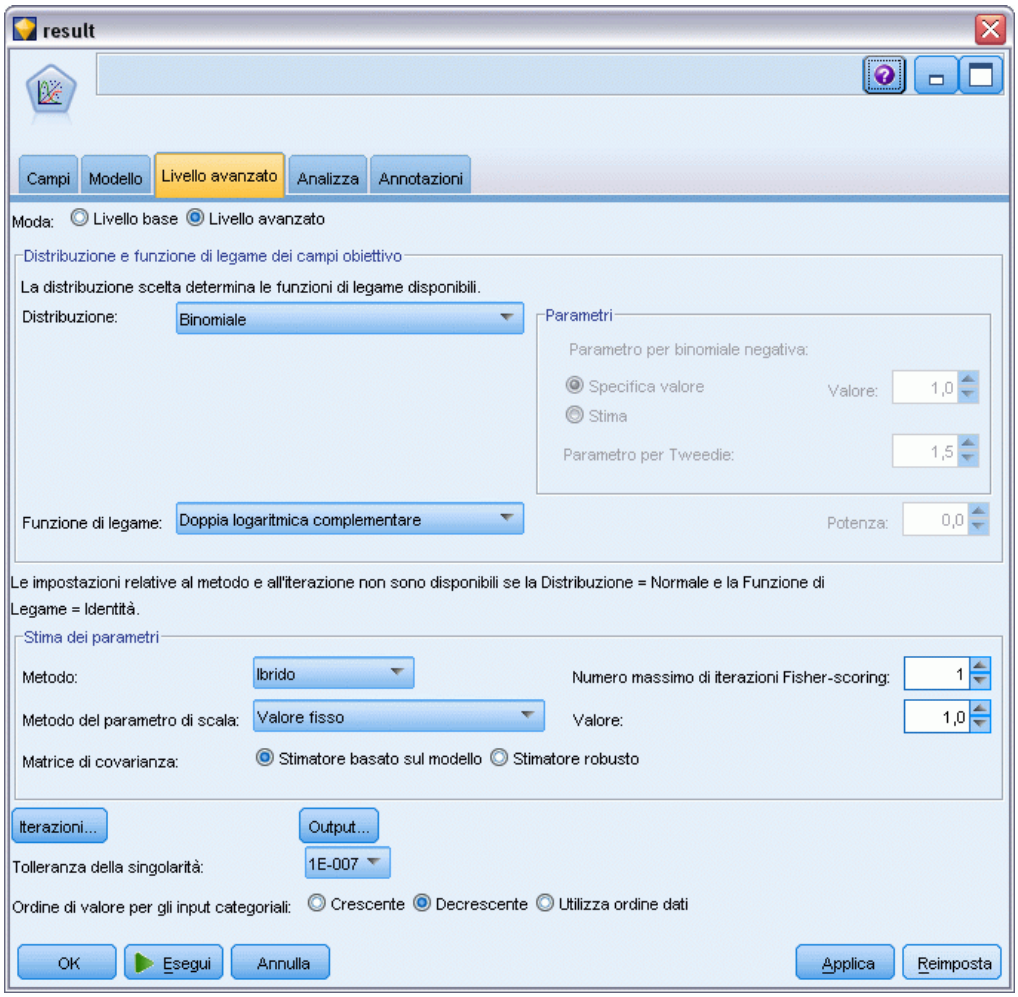


▶ Selezionare Prima (più bassa) come categoria di riferimento per l'obiettivo. Ciò indica che la seconda categoria è l'evento di interesse e che i suoi effetti sul modello riguardano l'interpretazione delle stime dei parametri.

▶ Deselezionare Includi l'intercetta nel modello.

- Fare clic sulla scheda Livello avanzato e selezionare Livello avanzato per attivare le opzioni di modellazione avanzata.

Figura 22-21
Scelta delle opzioni Expert



- Selezionare Binomiale come distribuzione e Doppia logaritmica complementare come funzione di legame.
- Selezionare Valore fisso come metodo per la stima del parametro di scala e lasciare il valore di default impostato su 1.0.
- Selezionare Decrescente come ordine delle categorie per i fattori. Questo indica che la prima categoria di ogni fattore sarà la relativa categoria di riferimento; l'effetto di questa selezione sul modello è nell'interpretazione delle stime dei parametri.
- Eseguire lo stream per creare l'insieme di modelli, che viene aggiunto allo stream e alla palette Modelli nell'angolo superiore destro. Per visualizzare i dettagli dei modelli, fare clic con il pulsante destro del mouse sull'insieme di modelli e selezionare Modifica o Sfoglia.

Test degli effetti del modello

Figura 22-22
Test degli effetti del modello per il modello effetti principali

Sorgente	Tipo III		
	Chi-quadrato di Wald	df	Sig.
period	,464	1	,496
duration	,000	1	,988
treatment	,117	1	,732
age	,314	1	,575
Variabile dipendente: Result by periodModello: period, duration, treatment, age			

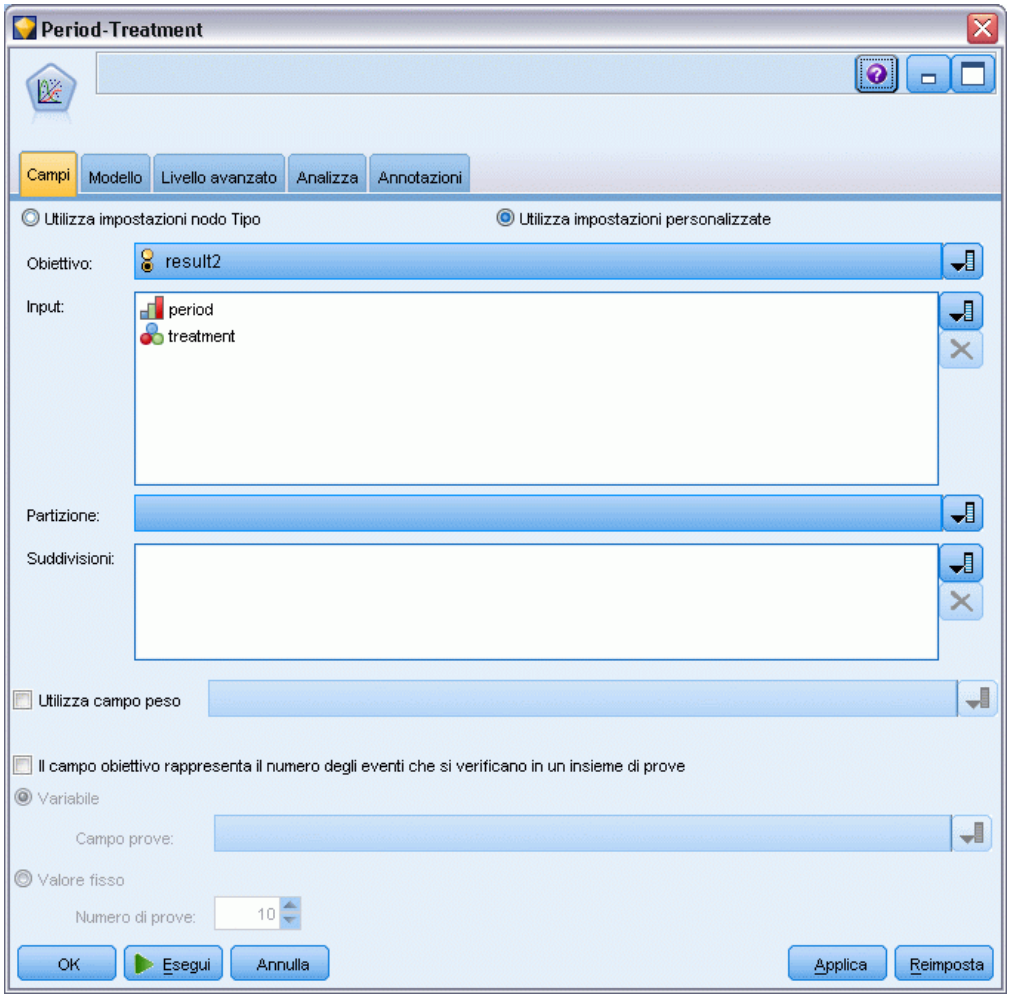
Nessuno degli effetti del modello è statisticamente significativo; tuttavia, qualsiasi differenza osservabile nel periodo e negli effetti del trattamento è di interesse clinico, per cui si adatterà un modello ridotto con questi termini soltanto.

Adattamento del modello ridotto

- Nella scheda Campi del nodo GenLin, fare clic su Utilizza impostazioni personalizzate.
- Selezionare *result2* come obiettivo.

- Selezionare *period* e *treatment* come input.

Figura 22-23
Scelta delle opzioni dei campi



- Eseguire il nodo e visualizzare il modello generato, quindi copiare il modello generato nella palette, collegare un nodo Tabella ed eseguirlo.

Stime di parametri

Figura 22-24
Stime parametri per il modello solo trattamento

Parametro	B	Deviazione standard Errore	95% Intervallo di confidenza di Wald		Test dell'ipotesi		
			Inferiore	Superiore	Chi-quadrato di Wald	df	Sig.
[period=2]	-1,794	,5792	-2,929	-,659	9,597	1	,002
[period=1]	-2,206	,5912	-3,365	-1,047	13,926	1	,000
[treatment=1]	,195	,6279	-1,035	1,426	,097	1	,756
[treatment=0]	0(a)	.	.	.	.	.	.
(Scala)	1(b)						

Variabile dipendente: Result by periodModello: period, treatment

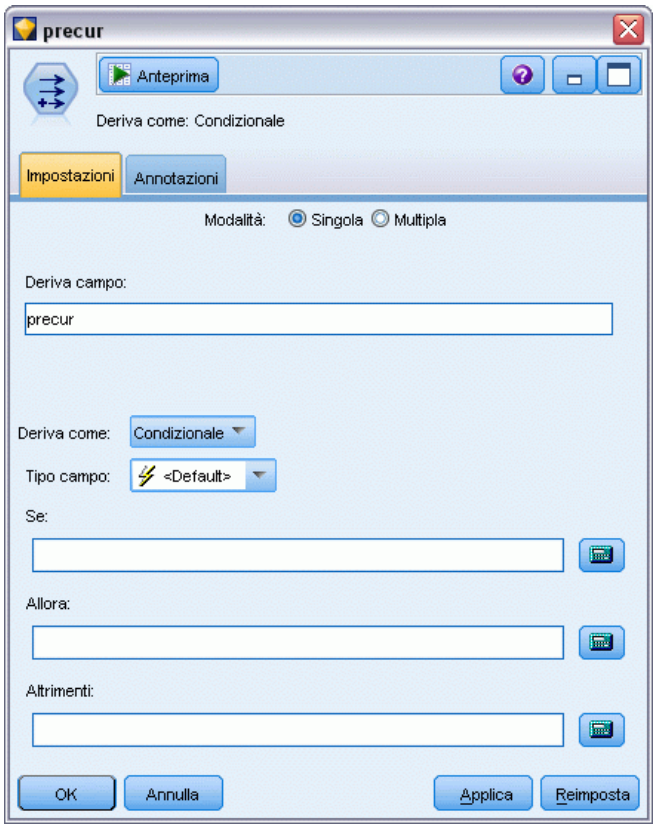
a. Impostato su zero poiché questo parametro è ridondante.

b. Fissato al valore visualizzato.

L'effetto del trattamento non è ancora statisticamente significativo, ma suggerisce solo che il trattamento *A* potrebbe essere migliore di *B* perché la stima dei parametri per il trattamento *B* è associata a una maggiore probabilità di recidiva nei primi 12 mesi. I valori del periodo, statisticamente significativi, sono diversi da 0, ma ciò avviene perché non è stato adattato un termine intercetta. L'effetto del periodo (la differenza tra i valori del predittore lineare per *[period=1]* e *[period=2]*) non è statisticamente significativo, come si può vedere nei test degli effetti del modello. Il predittore lineare (effetto del periodo + effetto del trattamento) è una stima di $\log(-\log(1-P(recur_{p,\,t}))$, dove $P(recur_{p,\,t})$ è la probabilità di recidiva al periodo p(=1 o 2, a indicare sei mesi o dodici mesi) dato il trattamento t(=*A* o *B*). Queste probabilità previste vengono generate per ogni osservazione nell'insieme di dati.

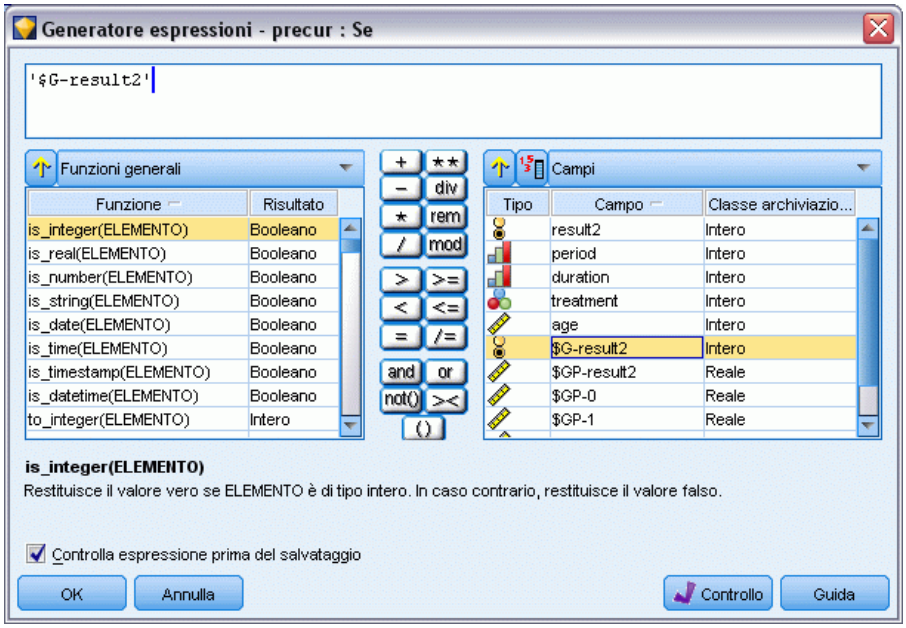
Recidive previste e probabilità di sopravvivenza

Figura 22-25
Opzioni della scheda Impostazioni per il nodo Nuovo campo



- Per ogni paziente, il modello calcola il punteggio del risultato previsto e la probabilità di tale risultato previsto. Per visualizzare le probabilità di recidiva previste, copiare il modello generato nella palette e collegare un nodo Nuovo campo.
- Nella scheda Impostazioni, immettere precur come nuovo campo.
- Scegliere la derivazione Condizionale.
- Fare clic sul pulsante del calcolatore per aprire il Generatore di espressioni per la condizione Se.

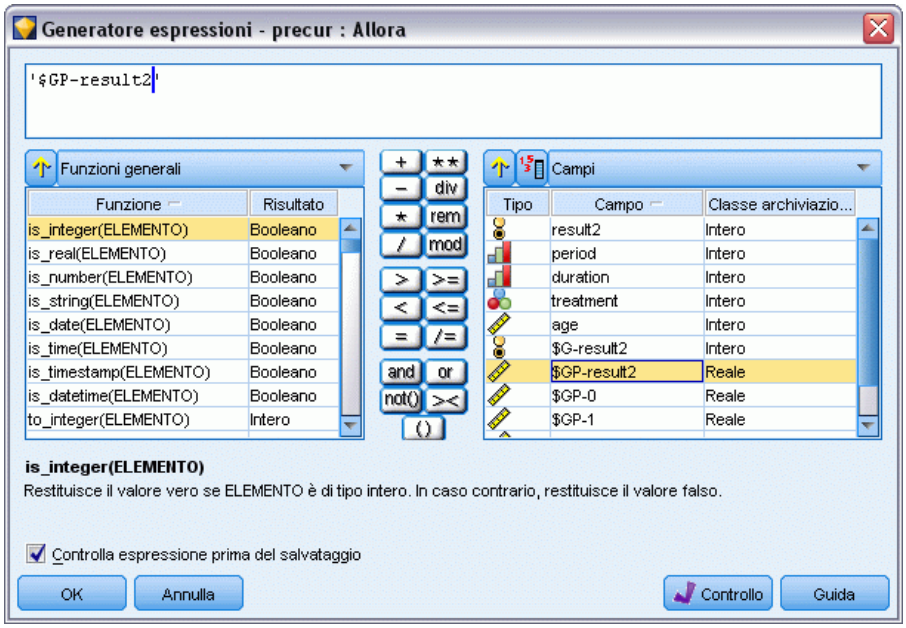
Figura 22-26
Nodo Nuovo campo: Generatore di espressioni per la condizione Se



- Inserire il campo *$G-risultato2* nell'espressione.
- Fare clic su OK.

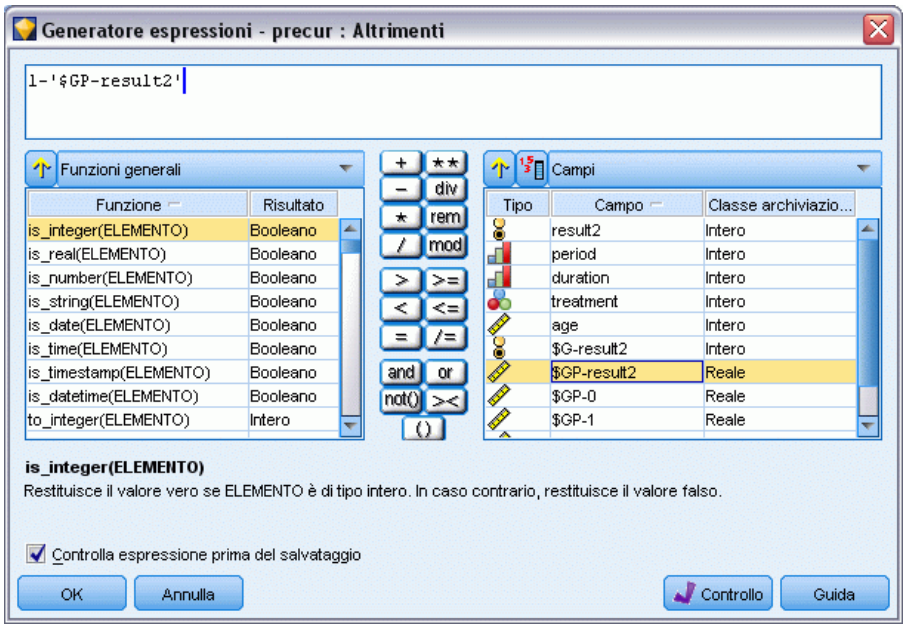
Il nuovo campo *precur* prende solo il valore dell'espressione Allora quando *$G-risultato2* è uguale a 1 e il valore dell'espressione Altrimenti quando è 0.

Figura 22-27
Nodo Nuovo campo: Generatore di espressioni per l'espressione Allora



- Fare clic sul pulsante del calcolatore per aprire il Generatore di espressioni per l'espressione Allora.
- Inserire il campo *$GP-risultato2* nell'espressione.
- Fare clic su OK.

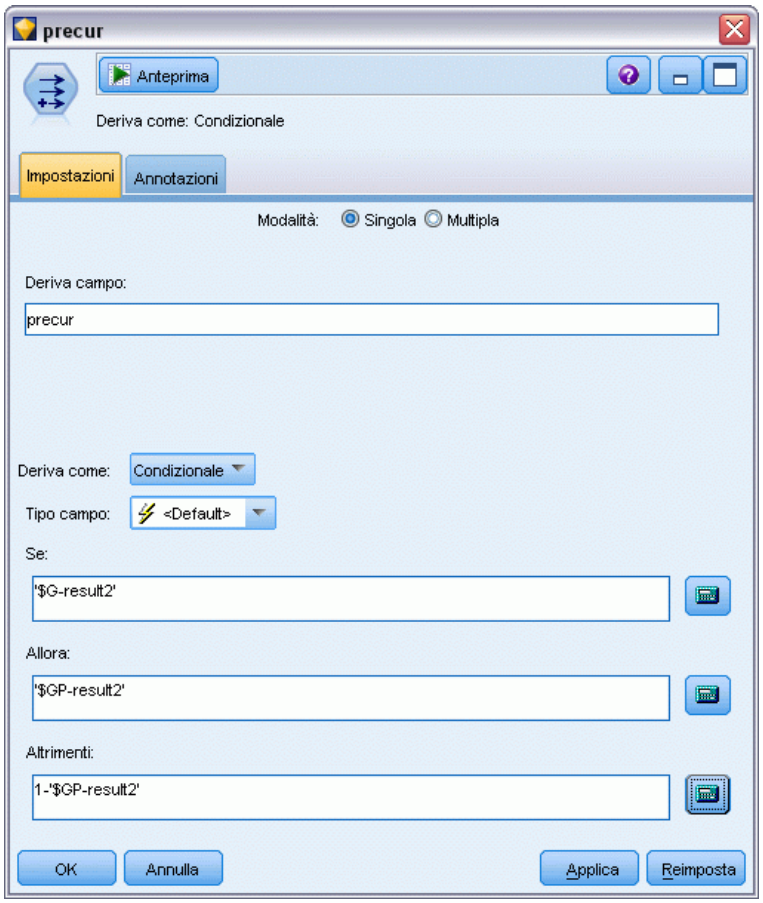
Figura 22-28
Nodo Nuovo campo: Generatore di espressioni per l'espressione Else



- Fare clic sul pulsante del calcolatore per aprire il Generatore di espressioni per l'espressione Else.

- Immettere 1- nell'espressione e quindi inserire il campo *$GP-risultato2* nell'espressione.
- Fare clic su OK.

Figura 22-29
Opzioni della scheda Impostazioni per il nodo Nuovo campo



- Collegare un nodo Tabella al nodo Nuovo campo e avviarne l'esecuzione.

Figura 22-30
Probabilità attese

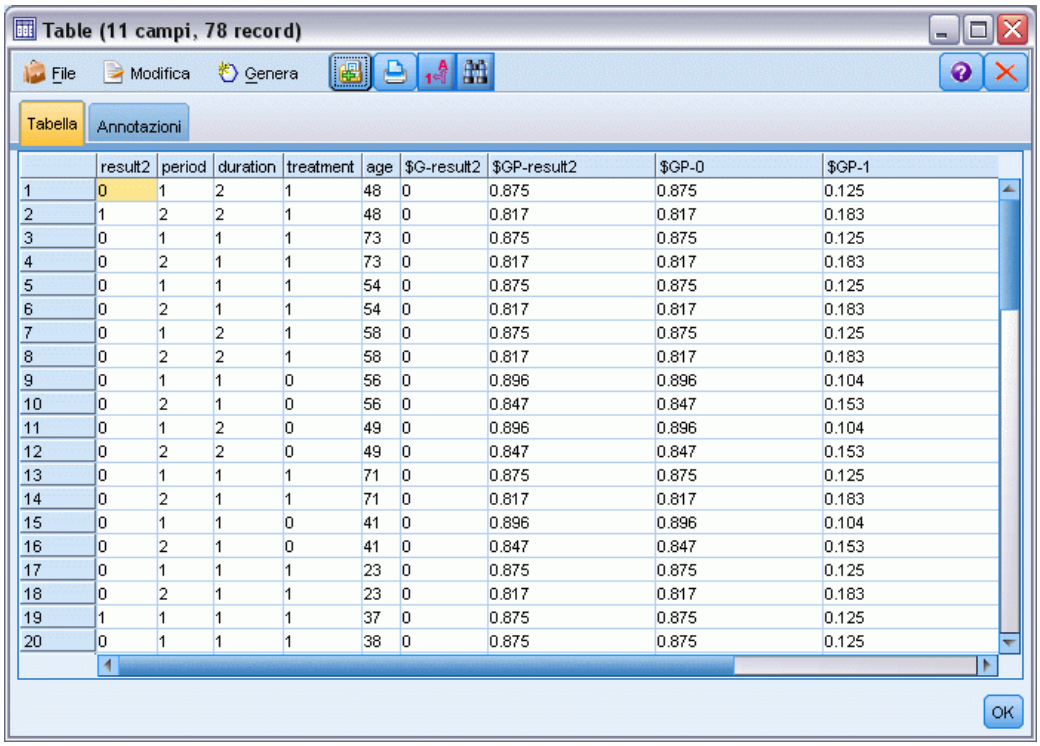


Table (11 campi, 78 record)

File Modifica Genera

Tabella Annotazioni

	result2	period	duration	treatment	age	$G-result2	$GP-result2	$GP-0	$GP-1
1	0	1	2	1	48	0	0.875	0.875	0.125
2	1	2	2	1	48	0	0.817	0.817	0.183
3	0	1	1	1	73	0	0.875	0.875	0.125
4	0	2	1	1	73	0	0.817	0.817	0.183
5	0	1	1	1	54	0	0.875	0.875	0.125
6	0	2	1	1	54	0	0.817	0.817	0.183
7	0	1	2	1	58	0	0.875	0.875	0.125
8	0	2	2	1	58	0	0.817	0.817	0.183
9	0	1	1	0	56	0	0.896	0.896	0.104
10	0	2	1	0	56	0	0.847	0.847	0.153
11	0	1	2	0	49	0	0.896	0.896	0.104
12	0	2	2	0	49	0	0.847	0.847	0.153
13	0	1	1	1	71	0	0.875	0.875	0.125
14	0	2	1	1	71	0	0.817	0.817	0.183
15	0	1	1	0	41	0	0.896	0.896	0.104
16	0	2	1	0	41	0	0.847	0.847	0.153
17	0	1	1	1	23	0	0.875	0.875	0.125
18	0	2	1	1	23	0	0.817	0.817	0.183
19	1	1	1	1	37	0	0.875	0.875	0.125
20	0	1	1	1	38	0	0.875	0.875	0.125

OK

È possibile riepilogare le probabilità di recidiva stimate come segue:

Trattamento	**6 mesi**	**12 mesi**
T	0.104	0.153
P	0.125	0.183

Da qui, le probabilità di sopravvivenza a 12 mesi possono essere stimate come $1-(P(recur_{1,t}) + P(recur_{2,t})\times(1-P(recur_{1,t})))$; pertanto, per ogni trattamento:

A: 1 − (0.104 + 0.153*0.896) = 0.759

B: 1 − (0.125 + 0.183*0.875) = 0.715

che mostra nuovamente un supporto non significativo dal punto di vista statistico per *A* come trattamento migliore.

Riepilogo

Utilizzando i Modelli lineari generalizzati, è stato possibile adattare una serie di modelli di regressione doppia logaritmica complementare a dati di sopravvivenza censurati per intervalli. Sebbene i dati facciano propendere a favore della scelta del trattamento *A*, il raggiungimento di

un risultato statisticamente significativo richiederebbe uno studio più vasto. Tuttavia, esistono ulteriori possibilità di esplorazione con i dati esistenti.

- Potrebbe essere utile riadattare il modello con effetti di interazione, in particolare tra *Periodo* e *Gruppo di trattamento*.

Per una spiegazione dei fondamenti matematici dei metodi di modellazione utilizzati in IBM® SPSS® Modeler vedere il manuale *SPSS Modeler Algorithms Guide*.

Capitolo
23

Uso della regressione di Poisson per analizzare la percentuale di danneggiamento delle navi (modelli lineari generalizzati)

Un modello lineare generalizzato può essere utilizzato per adattare una regressione di Poisson all'analisi dei dati dei conteggi. Ad esempio, un insieme di dati presentato e analizzato altrove riguarda i danni subiti dalle navi da carico a causa delle onde. I conteggi degli incidenti possono essere presentati in modelli con un tasso di Poisson in base ai valori dei predittori e il modello risultante può aiutare a determinare quali tipi di navi sono più soggetti a subire danni.

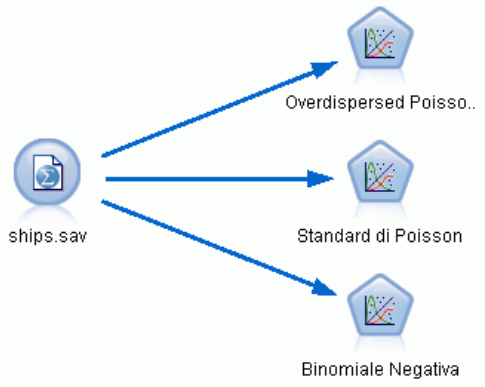
Questo esempio utilizza il flusso chiamato *ships_genlin.str*, che fa riferimento al file di dati *ships.sav*. Il file di dati si trova nella cartella *Demos* e il file di stream nella sottocartella *streams*. Per ulteriori informazioni, vedere l'argomento Cartella Demos in il capitolo 1 in *Manuale dell'utente di IBM SPSS Modeler 15*.

La creazione di modelli relativi ai conteggi delle celle può essere fuorviante in questa situazione poiché *Aggregate months of service* (Mesi di servizio totali) varia in base al tipo di nave. Variabili tipo questa, che misurano la percentuale di "esposizione" al rischio, vengono gestite nel modello lineare generalizzato come variabili offset. Inoltre, in una regressione di Poisson si presume che il log della variabile dipendente sia lineare nei predittori. Quindi, per utilizzare i modelli lineari generalizzati per adattare una regressione di Poisson alla percentuale di incidenti, è necessario utilizzare il *Logarithm of aggregate months of service* (Logaritmo dei mesi di servizio totali).

Adattamento di una regressione di Poisson "sovradispersa"

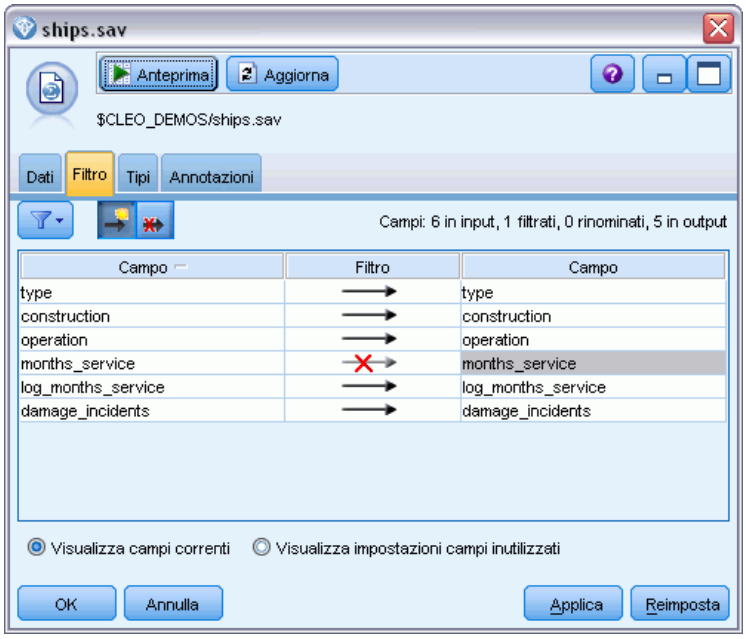
- Aggiungere un nodo di input File Statistics che punta a *ships.sav* nella cartella *Demos*.

Figura 23-1
Flusso di esempio per l'analisi della percentuale di danneggiamento



- Nella scheda Filtro del nodo sorgente, escludere il campo *months_service*. I valori di questa variabile trasformati tramite logaritmo sono contenuti in *log_months_service*, il cui utilizzo avverrà nell'analisi.

Figura 23-2
Filtraggio di un campo non necessario

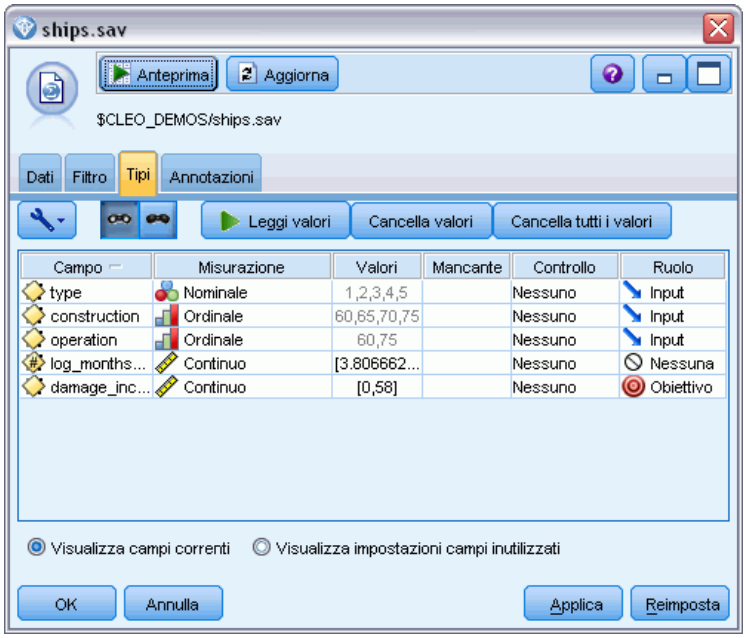


(In alternativa, nella scheda Tipi è possibile modificare il ruolo di questo campo in Nessuno, anziché escluderlo, oppure selezionare i campi da utilizzare nel nodo Modelli).

- Nella scheda Tipi del nodo di input, impostare il ruolo del campo *damage_incidents* su Obiettivo. Tutti gli altri campi dovrebbero avere il ruolo impostato su Input.

▶ Fare clic su Read Values (Leggi i valori) per creare un'istanza dei dati.

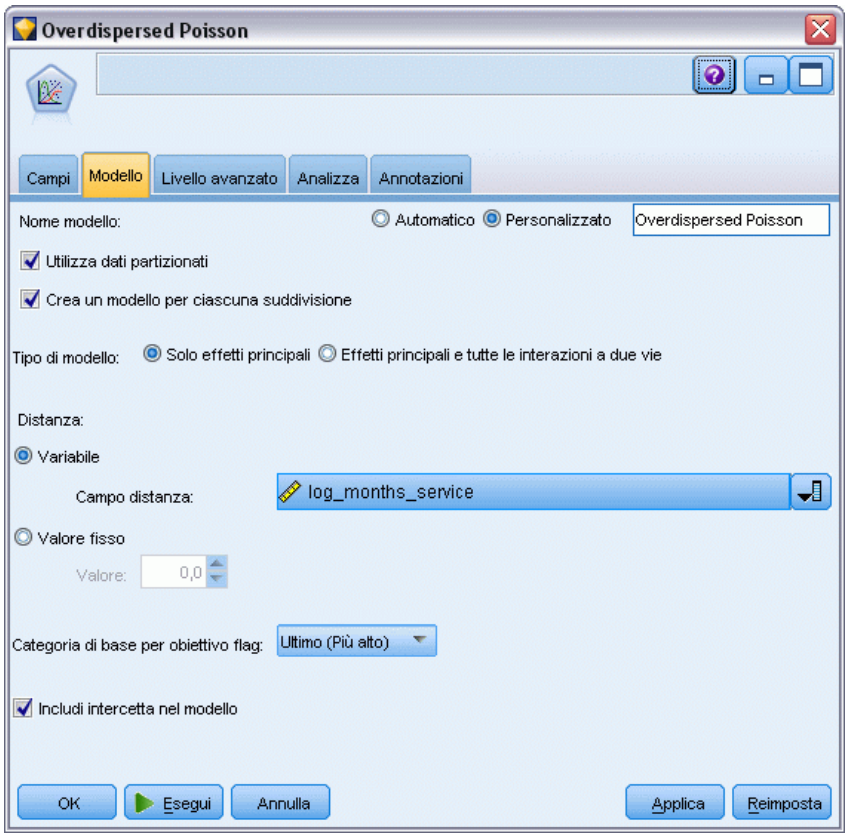
Figura 23-3
Impostazione del ruolo del campo



▶ Collegare un nodo Genlin al nodo di input; nel nodo Genlin, fare clic sulla scheda Modello.

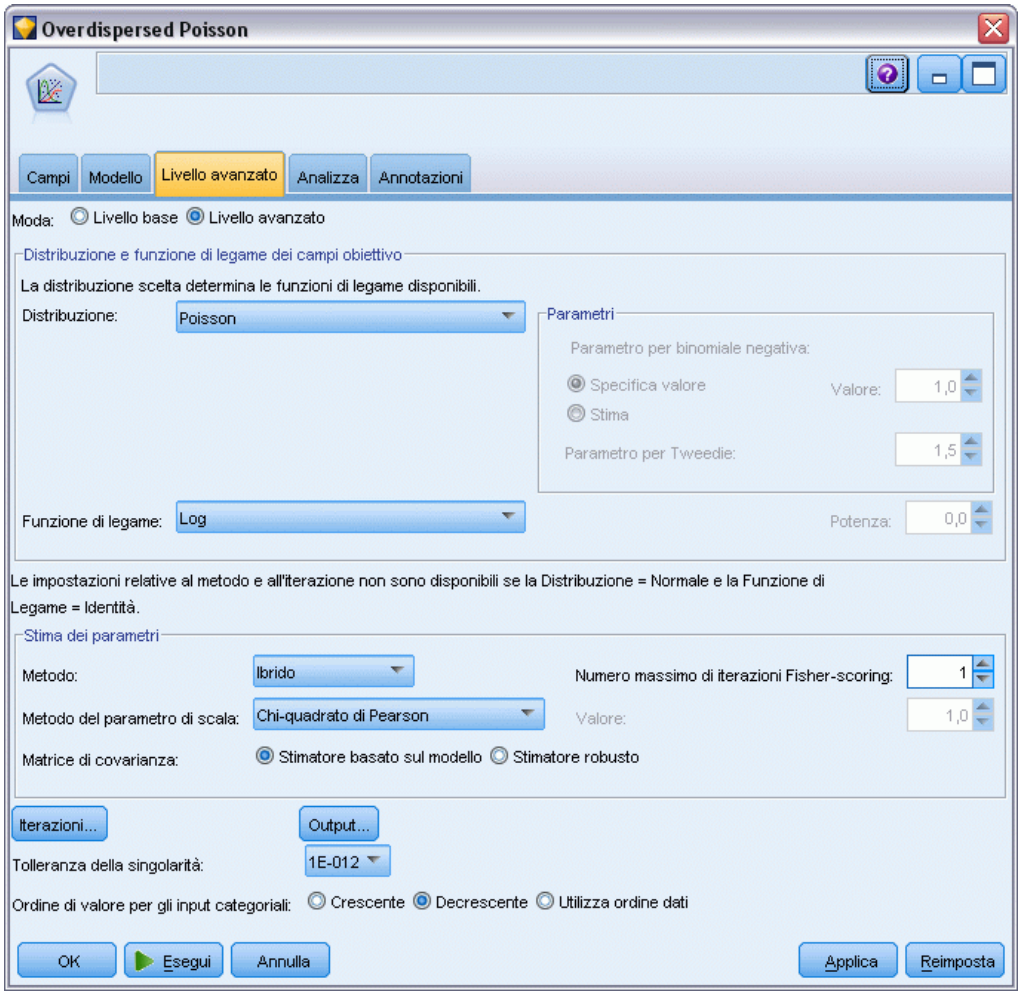
- Selezionare *log_months_service* come variabile di offset.

Figura 23-4
Scelta delle opzioni del modello



▶ Fare clic sulla scheda Livello avanzato e selezionare Livello avanzato per attivare le opzioni di modellazione avanzata.

Figura 23-5
Scelta delle opzioni Expert



▶ Selezionare Poisson come distribuzione per la risposta e Log come funzione di collegamento.

▶ SelezionareChi-quadrato di Pearson come metodo di stima del parametro di scala. In una regressione di Poisson, per il valore del parametro di scala viene solitamente utilizzato 1 ma McCullagh e Nelder utilizzano la stima del chi-quadrato di Pearson per ottenere stime di varianza e livelli di significatività più conservativi.

▶ Selezionare Decrescente come ordine delle categorie per i fattori. Questo indica che la prima categoria di ogni fattore sarà la relativa categoria di riferimento; l'effetto di questa selezione sul modello è nell'interpretazione delle stime dei parametri.

▶ Fare clic su Esegui per creare l'insieme di modelli, che viene aggiunto all'area di disegno dello stream e alla palette Modelli nell'angolo superiore destro. Per visualizzare i dettagli del modello,

fare clic con il pulsante destro del mouse sull'insieme di modelli e scegliere Modifica o Sfoglia, quindi selezionare la scheda Opzioni avanzate.

Statistiche di bontà dell'adattamento

Figura 23-6
Statistiche sulla bontà dell'adattamento

	Valore	df	Valore/df
Devianza	38,695	25	1,548
Devianza scala	38,695	25	
Chi-quadrato Pearson	42,275	25	1,691
Chi-quadrato scalato di Pearson	42,275	25	
Verosimiglianza(b)	-68,281		
Criterio di informazione di Akaike (AIC)	154,562		
AIC campione finito corretto (AICC)	162,062		
Criterio informativo di Bayes (BIC)	168,299		
AIC coerente (CAIC)	177,299		
Variabile dipendente: Number of damage incidentsModello: (Intercetta), type, construction, operation, offset = log_months_service			
a. I criteri di informazione sono in forma piccola migliore			
b. La funzione della verosimiglianza completa viene visualizzata e utilizzata nel calcolo dei criteri di informazione.			

La tabella delle statistiche sulla bontà di adattamento fornisce misure utili per il confronto di modelli concorrenti. Inoltre, il *valore/df* della devianza e delle statistiche chi-quadrato di Pearson fornisce stime corrispondenti per il parametro di scala. Questi valori devono essere prossimi a 1 per una regressione di Poisson; il fatto che tali valori siano maggiori di 1 indica che l'adattamento del modello sovradisperso potrebbe essere ragionevole.

Test omnibus

Figura 23-7
Test omnibus

Chi-quadrato per il rapporto di verosimiglianza	df	Sig.
107,633	8	,000
Variabile dipendente: Number of damage incidentsModello: (Intercetta), type, construction, operation, offset = log_months_service		
a. Confronta il modello adattato con il modello con la sola intercetta		

Il test omnibus è un test del rapporto di verosimiglianza chi-quadrato del modello corrente rispetto al modello null (in questo caso, intercetta). Il valore di significatività inferiore a 0,05 indica che il modello corrente è migliore del modello null.

Test degli effetti del modello

Figura 23-8
Test degli effetti del modello

Sorgente	Tipo III		
	Chi-quadrato di Wald	df	Sig.
(Intercetta)	2138,657	1	,000
type	15,415	4	,004
construction	17,242	3	,001
operation	6,249	1	,012
Variabile dipendente: Number of damage incidentsModello: (Intercetta), type, construction, operation, offset = log_months_service			

Ogni termine nel modello è testato per verificarne gli eventuali effetti. I termini con valori di significatività inferiori a 0,05 hanno un effetto rilevabile. Ogni termine di effetti principali contribuisce al modello.

Stime di parametri

Figura 23-9
Stime dei parametri

Parametro	B	Deviazione standard Errore	95% Intervallo di confidenza di Wald		Test dell'ipotesi		
			Inferiore	Superiore	Chi-quadrato di Wald	df	Sig.
(Intercetta)	-6,406	,2828	-6,960	-5,852	513,238	1	,000
[type=5]	,326	,3067	-,276	,927	1,127	1	,288
[type=4]	-,076	,3779	-,817	,665	,040	1	,841
[type=3]	-,687	,4279	-1,526	,151	2,581	1	,108
[type=2]	-,543	,2309	-,996	-,091	5,536	1	,019
[type=1]	0(a)	.	.	.	.	.	.
[construction=75]	,453	,3032	-,141	1,048	2,236	1	,135
[construction=70]	,818	,2208	,386	1,251	13,743	1	,000
[construction=65]	,697	,1946	,316	1,079	12,835	1	,000
[construction=60]	0(a)	.	.	.	.	.	.
[operation=75]	,384	,1538	,083	,686	6,249	1	,012
[operation=60]	0(a)	.	.	.	.	.	.
(Scala)	1,691(b)						

Variabile dipendente: Number of damage incidentsModello: (Intercetta), type, construction, operation, offset = log_months_service

a. Impostato su zero poiché questo parametro è ridondante.

b. Calcolo basato sul chi-quadrato di Pearson.

La tabella delle stime dei parametri riepiloga l'effetto di ogni predittore. Se l'interpretazione dei coefficienti in questo modello è difficile a causa della natura della funzione di collegamento, i segni dei coefficienti per le covariate e i valori relativi dei coefficienti per livelli di fattore rivelano molto bene gli effetti dei predittori nel modello.

- Relativamente alle covariate, i coefficienti positivi (negativi) indicato relazioni positive (inverse) tra predittori e risultato. Un valore crescente di una covariata con un coefficiente positivo corrisponde a una percentuale crescente degli incidenti che causano danni.
- Per i fattori, un livello di fattore con un coefficiente maggiore indica una maggiore incidenza di danneggiamento. Il segno di un coefficiente per un livello di fattore dipende dall'effetto del livello di fattore relativo alla categoria di riferimento.

È possibile elaborare le interpretazioni seguenti in base alle stime dei parametri:

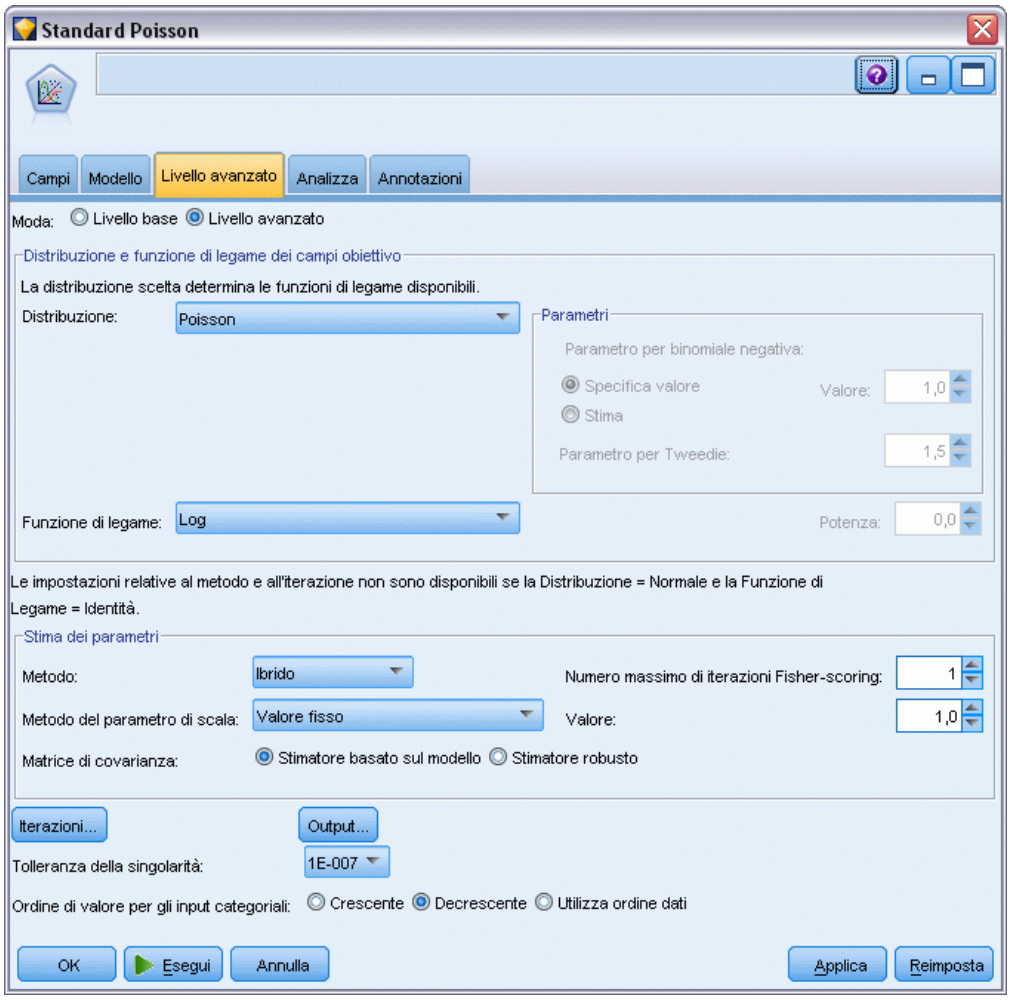
- Il tipo di nave *B [type=2]* ha una percentuale di danneggiamento statisticamente (valore*p*di 0,019) inferiore (coefficiente stimato di –0,543) rispetto al tipo di nave *A [type=1]*, ovvero la categoria di riferimento. Il tipo *C [type=3]* ha un parametro stimato inferiore a *B*, ma la variabilità nella stima *C* offusca l'effetto. Vedere le medie marginali stimate per tutte le relazioni tra i livelli dei fattori.

- Le navi costruite tra il 1965–69 *[construction=65]* e tra il 1970–74 *[construction=70]* hanno statisticamente (valori*p* <0,001) percentuali di danneggiamento superiori (coefficienti stimati di 0,697 e 0,818, rispettivamente) rispetto a quelle costruite tra il 1960–64 *[construction=60]*, ovvero la categoria di riferimento. Vedere le medie marginali stimate per tutte le relazioni tra i livelli dei fattori.
- Le navi in attività tra il 1975–79 *[operation=75]* hanno statisticamente (valore*p* di 0,012) una percentuale di danneggiamento superiore (coefficiente stimato di 0,384) rispetto a quelle in attività tra il 1960–1974 *[operation=60]*.

Adattamento dei modelli alternativi

Un problema con la regressione di Poisson "sovradispersa" è dovuto al fatto che non esiste un modo formale per testarla rispetto alla regressione di Poisson "standard". Tuttavia, un test formale suggerito per determinare se esiste sovradispersione consiste nell'esecuzione di un test del rapporto di verosomiglianza tra una regressione di Poisson "standard" e una regressione negativa binomiale, mantenendo invariate tutte le altre impostazioni. Se nella regressione di Poisson non vi è sovradispersione, la statistica −2×(log-verosimiglianza per il modello di Poisson − log-verosimiglianza per il modello binomiale negativo) deve avere una distribuzione mista con metà della massa di probabilità a 0 e il resto in una distribuzione chi-quadrato con 1 grado di libertà.

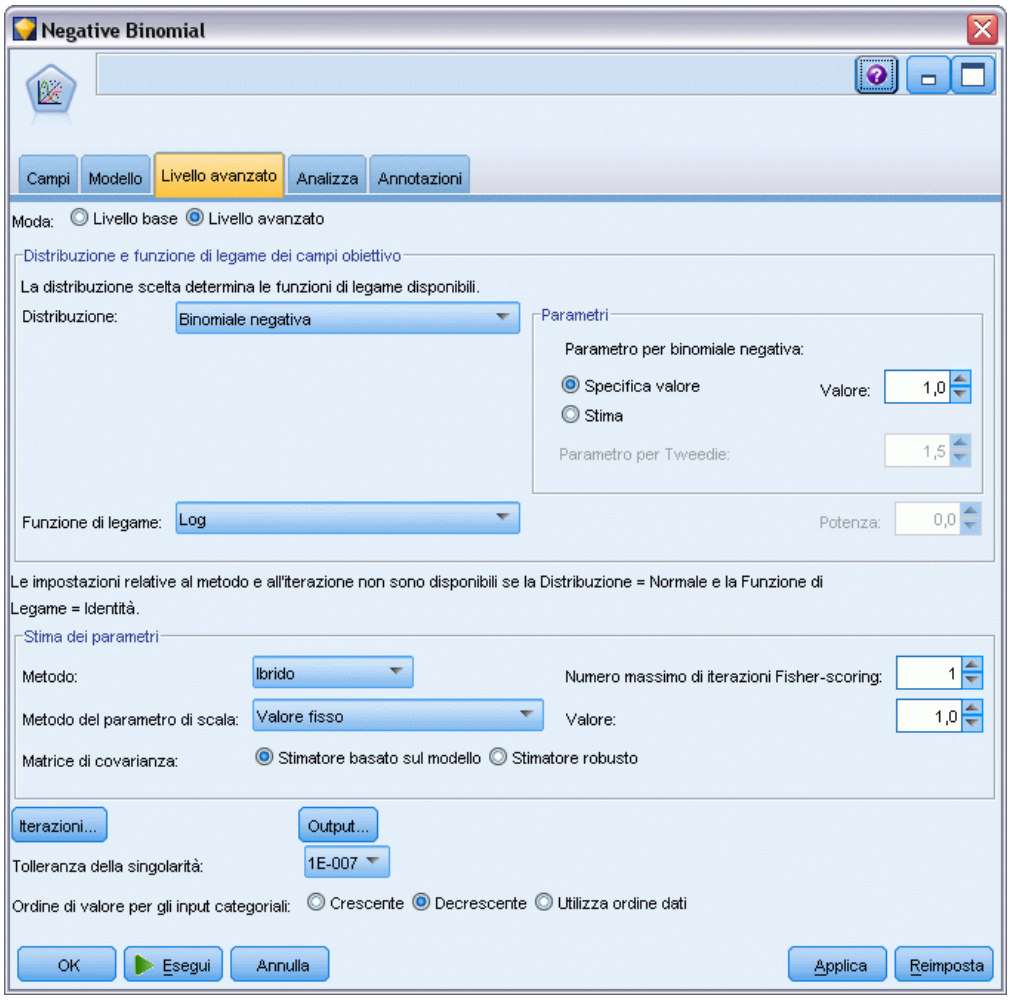
Figura 23-10
Scheda Expert



Per adattare la regressione di Poisson “standard”, copiare e incollare il nodo Genlin, allegarlo al nodo di input, aprire il nuovo nodo e fare clic sulla scheda Livello avanzato.

- Selezionare Valore fisso come metodo di stima del parametro di scala. Per impostazione predefinita, questo valore è 1.

Figura 23-11
Scheda Expert



- Per adattare la regressione binomiale negativa, copiare e incollare il nodo Genlin, allegarlo al nodo di input, aprire il nuovo nodo e fare clic sulla scheda Livello avanzato.
- Selezionare Binomiale negativo come distribuzione. Lasciare il valore predefinito 1 per il parametro ausiliario.
- Eseguire lo stream e verificare nella scheda Opzioni avanzate la presenza degli insiemi di modelli appena creati.

Statistiche di bontà dell'adattamento

Figura 23-12
Statistiche di bontà per la regressione di Poisson standard

	Valore	df	Valore/df
Devianza	38,695	25	1,548
Devianza scala	22,883	25	
Chi-quadrato Pearson	42,275	25	1,691
Chi-quadrato scalato di Pearson	25,000	25	
Verosimiglianza(b,c)	-68,281		
Criterio di informazione di Akaike (AIC)	154,562		
AIC campione finito corretto (AICC)	162,062		
Criterio informativo di Bayes (BIC)	168,299		
AIC coerente (CAIC)	177,299		
Variabile dipendente: Number of damage incidentsModello: (Intercetta), type, construction, operation, offset = log_months_service			
a. I criteri di informazione sono in forma piccola migliore			
b. La funzione della verosimiglianza completa viene visualizzata e utilizzata nel calcolo dei criteri di informazione.			

Il log verosomiglianza indicato per la regressione di Poisson standard è –68,281. Confrontarlo con il modello negativo binomiale.

Figura 23-13
Statistiche di bontà per la regressione negativa binomiale

	Valore	df	Valore/df
Devianza	11,145	25	,446
Devianza scala	11,145	25	
Chi-quadrato Pearson	8,815	25	,353
Chi-quadrato scalato di Pearson	8,815	25	
Verosimiglianza(b)	-83,725		
Criterio di informazione di Akaike (AIC)	185,450		
AIC campione finito corretto (AICC)	192,950		
Criterio informativo di Bayes (BIC)	199,187		
AIC coerente (CAIC)	208,187		
Variabile dipendente: Number of damage incidentsModello: (Intercetta), type, construction, operation, offset = log_months_service			
a. I criteri di informazione sono in forma piccola migliore			
b. La funzione della verosimiglianza completa viene visualizzata e utilizzata nel calcolo dei criteri di informazione.			

Il log verosomiglianza indicato per la regressione negativa binomiale è –83,725. Questo valore è di fatto *inferiore* al log verosomiglianza della regressione di Poisson che indica, senza dover eseguire un test del rapporto di verosomiglianza, che tale regressione negativa binomiale non offre un miglioramento alla regressione di Poisson.

Tuttavia, il valore scelto 1 per il parametro ausiliario della distribuzione negativa binomiale potrebbe non essere ottimale per questo insieme di dati. Un altro modo che è possibile testare per la sovradispersione è impostare un modello negativo binomiale con un parametro ausiliario su zero e richiedere il test del moltiplicatore di Lagrange nella finestra Output, scheda Expert. Se il test non è significativo, la sovradispersione non dovrebbe essere un problema per questo insieme di dati.

Riepilogo

Utilizzando i modelli lineari generalizzati, sono stati adattati tre differenti ai dati dei conteggi. Si è visto che la regressione binomiale negativa non offre alcun miglioramento rispetto alla regressione di Poisson. La regressione di Poisson sovradispersa sembra offrire un'alternativa accettabile al modello di Poisson standard ma non esiste un test formale per la scelta di uno di essi.

Per una spiegazione dei fondamenti matematici dei metodi di modellazione utilizzati in IBM® SPSS® Modeler vedere il manuale *SPSS Modeler Algorithms Guide*.

Capitolo
24

Adattamento della regressione gamma alle richieste di risarcimento a una compagnia di assicurazioni auto (modelli lineari generalizzati)

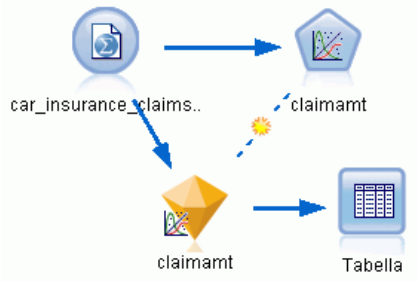
Un modello lineare generalizzato può essere utilizzato per adattare una regressione gamma all'analisi dei dati dell'intervallo positivo. Ad esempio, un insieme di dati presentato e analizzato altrove riguarda le richieste di risarcimento auto. La quantità media di richieste di risarcimento può essere adattata come avente una distribuzione gamma, utilizzando una funzione di collegamento inverso per correlare la media della variabile dipendente a una combinazione lineare di predittori. Per tener conto del numero differente di richieste di risarcimento utilizzate per calcolare le quantità medie di richieste di risarcimento, specificare *Numero di richieste di risarcimento* come peso scalato.

Questo esempio utilizza il flusso chiamato *car-insurance_genlin.str*, che fa riferimento al file di dati *car_insurance_claims.sav*. Il file di dati si trova nella cartella *Demos* e il file di stream nella sottocartella *streams*. Per ulteriori informazioni, vedere l'argomento Cartella Demos in il capitolo 1 in *Manuale dell'utente di IBM SPSS Modeler 15*.

Creazione del flusso

- Aggiungere un nodo di input File Statistics che punta a *car_insurance_claims.sav* nella cartella *Demos*.

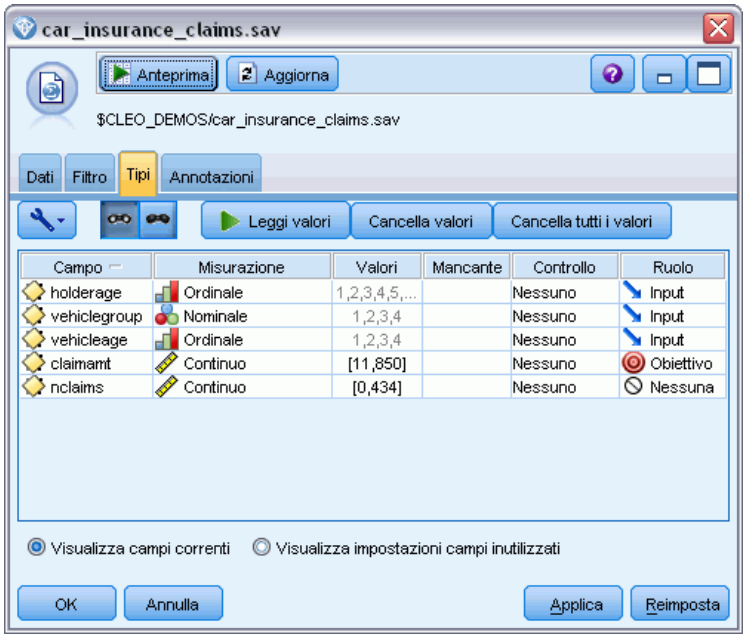
Figura 24-1
Flusso di esempio per stimare le richieste di risarcimento danni



- Nella scheda Tipi del nodo di input, impostare il ruolo del campo *claimamt* (qtàrichieste) su Obiettivo. Tutti gli altri campi dovrebbero avere il ruolo impostato su Input.

▶ Fare clic su Read Values (Leggi i valori) per creare un'istanza dei dati.

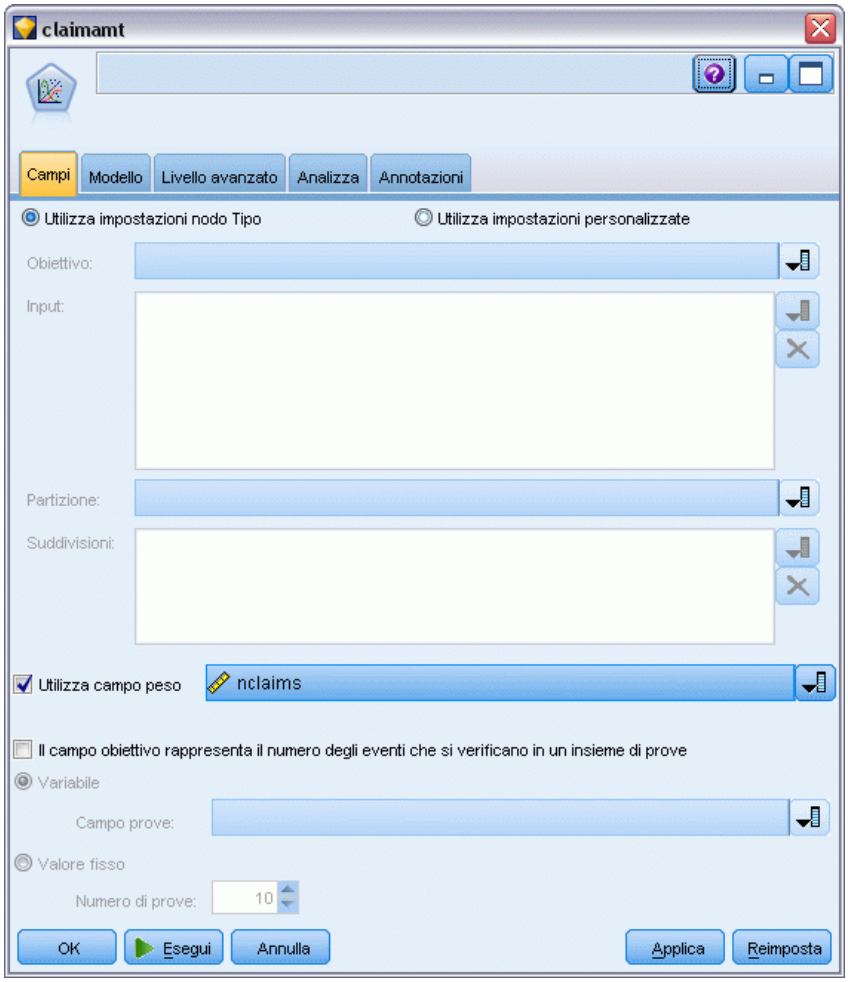
Figura 24-2
Impostazione del ruolo del campo



▶ Collegare un nodo Genlin al nodo di input; nel nodo Genlin, fare clic sulla scheda Campi.

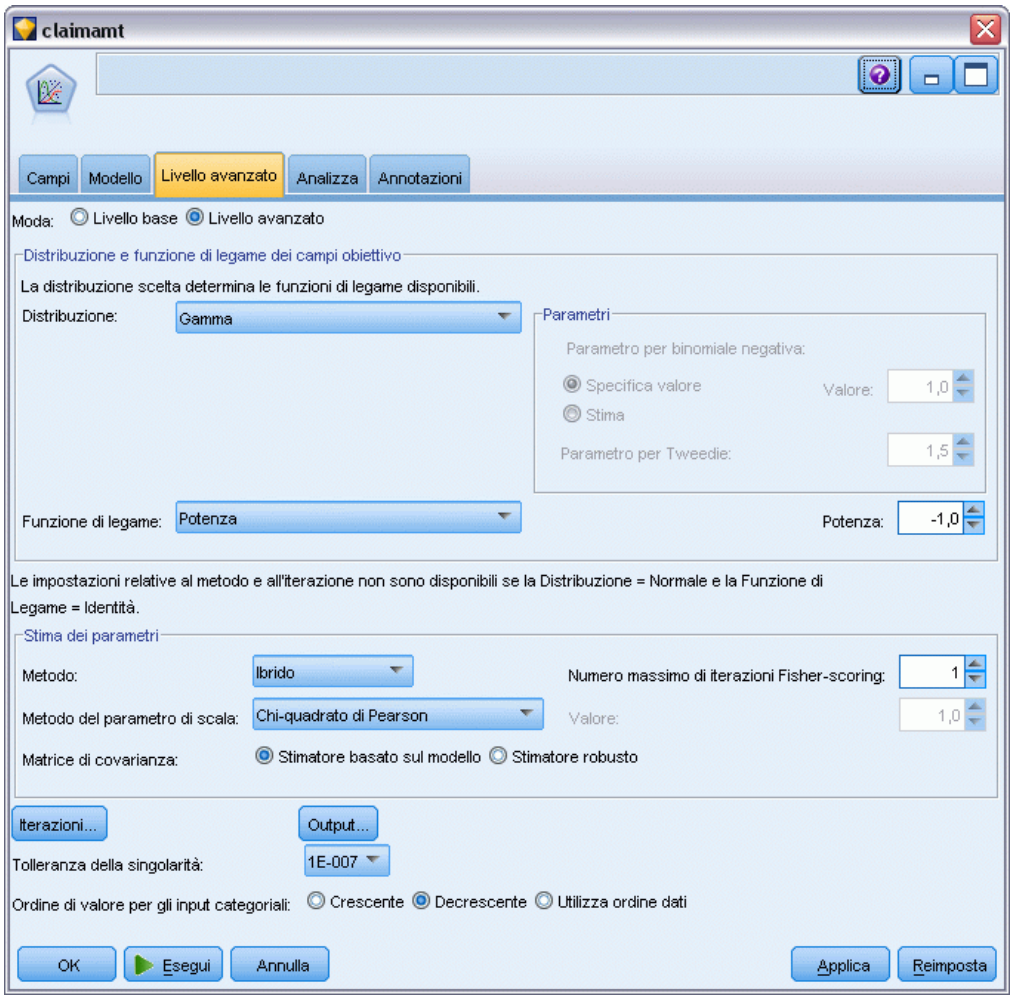
▶ Selezionare *Modelli lineari generalizzati* come campo di peso della scala.

Figura 24-3
Scelta delle opzioni dei campi



- Fare clic sulla scheda Expert e selezionare Expert per attivare le opzioni di creazione dei modelli per esperti.

Figura 24-4
Scelta delle opzioni Expert



- Selezionare Gamma come distribuzione della risposta.
- Selezionare Potenza come funzione di collegamento e digitare -1.0 come esponente della funzione di potenza. Questo è un collegamento inverso.
- SelezionareChi-quadrato di Pearson come metodo di stima del parametro di scala. Questo è il metodo utilizzato da McCullagh e Nelder e pertanto verrà utilizzato in questo esempio per replicarne i risultati.
- Selezionare Decrescente come ordine delle categorie per i fattori. Questo indica che la prima categoria di ogni fattore sarà la relativa categoria di riferimento; l'effetto di questa selezione sul modello è nell'interpretazione delle stime dei parametri.
- Fare clic su Esegui per creare l'insieme di modelli, che viene aggiunto allo stream e alla palette Modelli nell'angolo superiore destro. Per visualizzare i dettagli del modello, fare clic con il

pulsante destro sull'insieme di modelli e scegliere Modifica o Sfoglia, quindi selezionare la scheda Opzioni avanzate.

Stime di parametri

Figura 24-5
Stime dei parametri

Parametro	B	Deviazione standard Errore	95% Intervallo di confidenza di Wald		Test dell'ipotesi		
			Inferiore	Superiore	Chi-quadrato di Wald	df	Sig.
(Intercetta)	,003	,0004	,003	,004	66,593	1	,000
[holderage=8]	,001	,0004	,000	,002	4,898	1	,027
[holderage=7]	,001	,0004	,000	,002	5,046	1	,025
[holderage=6]	,001	,0004	,000	,002	5,740	1	,017
[holderage=5]	,001	,0004	,001	,002	10,682	1	,001
[holderage=4]	,000	,0004	,000	,001	1,268	1	,260
[holderage=3]	,000	,0004	,000	,001	,720	1	,396
[holderage=2]	,000	,0004	-,001	,001	,054	1	,816
[holderage=1]	0(a)	.	.	.	.	.	.
[vehiclegroup=4]	-,001	,0002	-,002	-,001	61,883	1	,000
[vehiclegroup=3]	-,001	,0002	-,001	,000	13,039	1	,000
[vehiclegroup=2]	3,77E-005	,0002	,000	,000	,050	1	,823
[vehiclegroup=1]	0(a)	.	.	.	.	.	.
[vehicleage=4]	,004	,0004	,003	,005	88,175	1	,000
[vehicleage=3]	,002	,0002	,001	,002	53,013	1	,000
[vehicleage=2]	,000	,0001	,000	,001	13,191	1	,000
[vehicleage=1]	0(a)	.	.	.	.	.	.
(Scala)	1,209(b)						

Variabile dipendente: Average cost of claimsModello: (Intercetta), holderage, vehiclegroup, vehicleage

a. Impostato su zero poiché questo parametro è ridondante.

b. Calcolo basato sul chi-quadrato di Pearson.

Il test omnibus e i test degli effetti del modello (non mostrati) indicano che il modello è migliore del modello null e che ogni termine dell'effetto principale contribuisce al modello. La tabella delle stime dei parametri mostra gli stessi valori ottenuti da McCullagh e Nelder per i livelli fattoriali e il parametro di scala .

Riepilogo

Utilizzando i modelli lineari generalizzati, è stata adattata una regressione gamma ai dati delle richieste di risarcimento. Tenere presente che sebbene sia stata utilizzata la funzione di collegamento canonica per la distribuzione gamma, anche un collegamento log fornisce risultati ragionevoli. In generale, è difficile se non impossibile confrontare direttamente i modelli con funzioni di collegamento differenti; tuttavia, il collegamento log è un caso speciale di collegamento power dove l'esponente è 0; quindi è possibile confrontare le devianze di un modello con un collegamento log e un modello con un collegamento power per determinare quale fornisce l'adattamento migliore (vedere, ad esempio, la sezione 11.3 di McCullagh e Nelder).

Per una spiegazione dei fondamenti matematici dei metodi di modellazione utilizzati in IBM® SPSS® Modeler vedere il manuale *SPSS Modeler Algorithms Guide*.

Capitolo
25

Classificazione dei campioni di cellule (SVM)

Support Vector Machine (SVM) è una tecnica di classificazione e regressione particolarmente indicata per gli insiemi di dati di grandi dimensioni. Per insieme di dati di grandi dimensioni si intende un insieme di dati con molti predittori, per esempio quelli che si trovano nel settore della bioinformatica (l'applicazione delle tecnologie dell'informazione ai dati biochimici e biologici).

Un ricercatore medico ha ricevuto un insieme di dati contenente le caratteristiche di numerosi campioni di cellule umane estratte da pazienti ritenuti a rischio di sviluppo di tumori. L'analisi dei dati originali ha dimostrato che molte caratteristiche erano sostanzialmente diverse a seconda dei campioni benigni o maligni. Il ricercatore desidera sviluppare un modello SVM in grado di utilizzare i valori di tali caratteristiche cellulari nei campioni di altri pazienti per fornire un'indicazione precoce della benignità o malignità di tali campioni.

In questo esempio viene utilizzato lo stream denominato *svm_cancer.str*, disponibile nella sottocartella *streams* della cartella *Demos*. Il file di dati è *cell_samples.data*. Per ulteriori informazioni, vedere l'argomento Cartella Demos in il capitolo 1 in *Manuale dell'utente di IBM SPSS Modeler 15*.

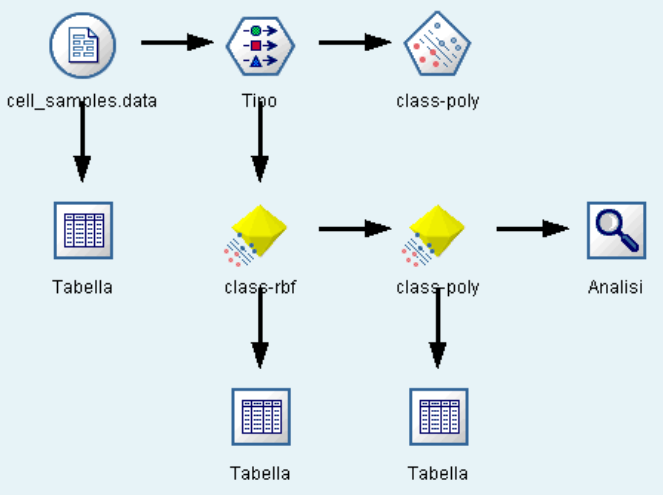
L'esempio si basa su un insieme di dati disponibile pubblicamente nel repository per l'apprendimento automatico UCI (Asuncion e Newman, 2007). L'insieme di dati è composto da diverse centinaia di record di campioni di cellule, ciascuno dei quali contiene i valori di un insieme di caratteristiche cellulari. I campi di ogni record sono:

Nome del campo	Descrizione
ID	Identificatore paziente
Agglutinazione	Spessore dell'agglutinazione
UnifDim	Uniformità della dimensione cellulare
UnifForma	Uniformità della forma cellulare
AdMarg	Adesione marginale
DimSingEp	Dimensione della singola cellula epiteliale
NucNudi	Nuclei nudi
CromBlan	Cromatina blanda
NuclNorm	Nucleolo normale
Mit	Mitosi
Class	Benigno o maligno

Ai fini di questo esempio, utilizziamo un insieme di dati con un numero di predittori relativamente ridotto in ogni record.

Creazione dello stream

Figura 25-1
Stream di esempio per mostrare la modellazione SVM



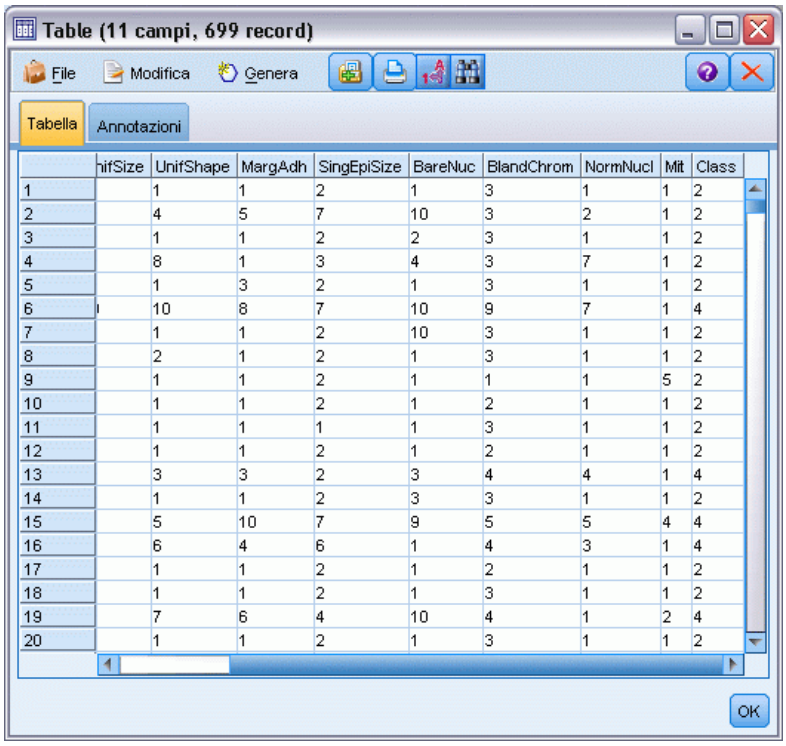
- Creare un nuovo stream e aggiungere un nodo di input Testo variabile che punti al file *cell_samples.data* situato nella directory *Demos* dell'installazione IBM® SPSS® Modeler.

  Osserviamo ora i dati nel file di origine.

- Aggiungere un nodo Tabella allo stream.

- Collegare il nodo Tabella al nodo Testo variabile ed eseguire lo stream.

Figura 25-2
Dati di origine per SVM

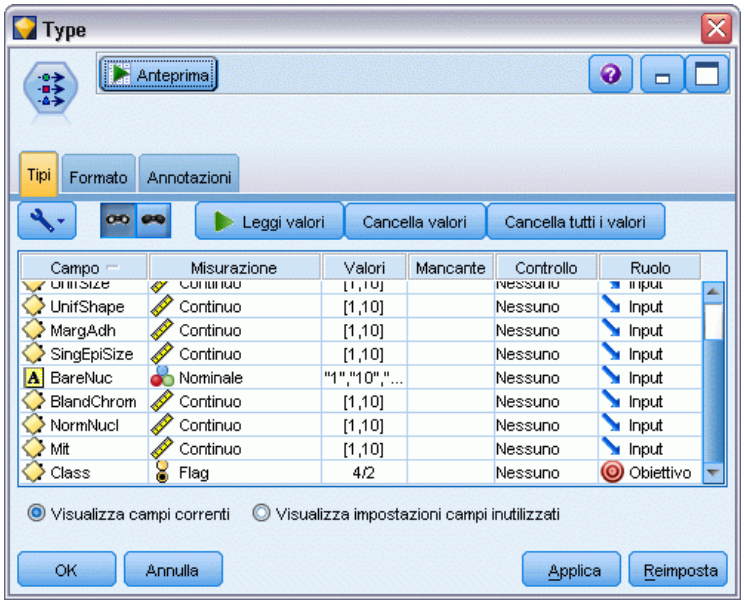


	hifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1		1	1	2	1	3	1	1	2
2		4	5	7	10	3	2	1	2
3		1	1	2	2	3	1	1	2
4		8	1	3	4	3	7	1	2
5		1	3	2	1	3	1	1	2
6		10	8	7	10	9	7	1	4
7		1	1	2	10	3	1	1	2
8		2	1	2	1	3	1	1	2
9		1	1	2	1	1	1	5	2
10		1	1	2	1	2	1	1	2
11		1	1	1	1	3	1	1	2
12		1	1	2	1	2	1	1	2
13		3	3	2	3	4	4	1	4
14		1	1	2	3	3	1	1	2
15		5	10	7	9	5	5	4	4
16		6	4	6	1	4	3	1	4
17		1	1	2	1	2	1	1	2
18		1	1	2	1	3	1	1	2
19		7	6	4	10	4	1	2	4
20		1	1	2	1	3	1	1	2

Il campo *ID* contiene gli identificatori dei pazienti. Le caratteristiche dei campioni di cellule di ogni paziente sono contenute nei campi da *Agglutinazione* a *Mit*. Questi valori sono classificati da 1 a 10, dove 1 è il valore più vicino alla benignità.

Il campo *Classe* contiene la diagnosi, confermata da procedure mediche separate, rispetto alla benignità (valore = 2) o malignità (valore = 4) dei campioni.

Figura 25-3
Impostazioni del nodo Tipo



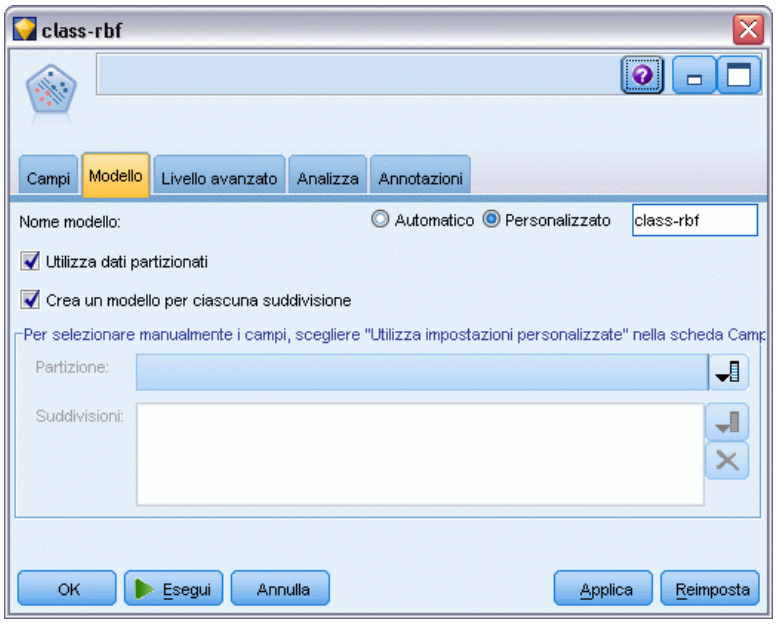
- Aggiungere un nodo Tipo e collegarlo al nodo Testo variabile.
- Aprire il nodo Tipo.

  L'obiettivo è che il modello preveda il valore della *Classe* [cioè, benigno (=2) o maligno (=4)]. Poiché questo campo può avere solo uno dei due valori possibili, è necessario modificare il livello di misurazione in modo da rispecchiare questa situazione.

- Nella colonna Misurazione del campo *Classe* (l'ultima dell'elenco), fare clic sul valore Continuo e impostarlo su Flag.
- Fare clic su Leggi valori.
- Nella colonna Ruolo, impostare il ruolo dell'*ID* (identificatore paziente) su Nessuno, in quanto questo valore non verrà utilizzato né come predittore né come obiettivo del modello.
- Impostare il ruolo dell'obiettivo, *Classe*, su Obiettivo e lasciare il ruolo di tutti gli altri campi (i predittori) impostato su Input.
- Fare clic su OK.

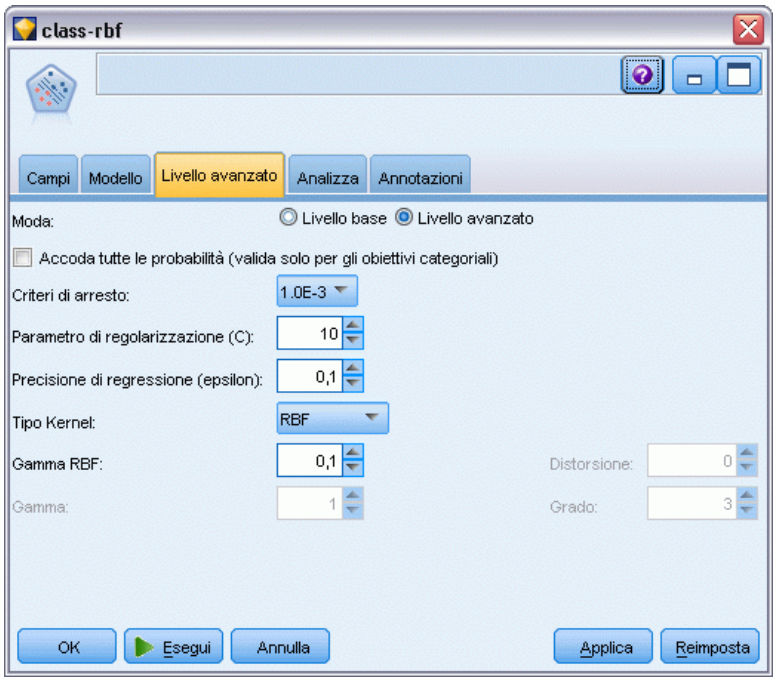
  Il nodo SVM offre una scelta di funzioni Kernel per eseguire la propria elaborazione. Poiché non esiste un modo semplice per conoscere quale funzione si adatti meglio a un determinato insieme di dati, sceglieremo varie funzioni a turno e ne confronteremo i risultati. Iniziamo con la funzione di default, RBF (Radial Basis Function).

Figura 25-4
Impostazioni della scheda Modello



- Dalla palette Modelli, collegare un nodo SVM al nodo Tipo.
- Aprire il nodo SVM. Nella scheda Modello, fare clic sull'opzione Personalizzato per Nome modello e immettere *classe-rbf* nel campo di testo adiacente

Figura 25-5
Impostazioni di default della scheda Livello avanzato



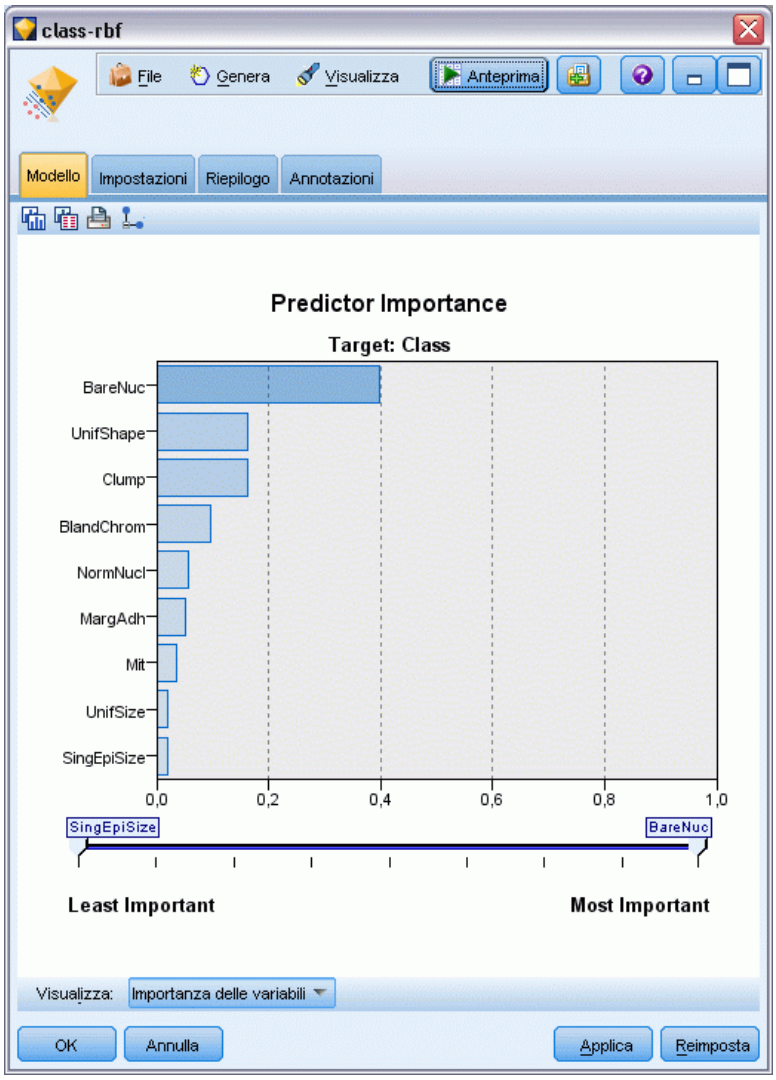
- Nella scheda Livello avanzato, impostare la Modalità su Livello avanzato per una migliore leggibilità ma lasciare invariate tutte le opzioni di default. Si noti che Tipo Kernel è impostato su RBF per default. Tutte le opzioni vengono visualizzate in grigio nella modalità Livello base.

Figura 25-6
Impostazioni della scheda Analisi

- Nella scheda Analisi, selezionare la casella di controllo Calcola importanza variabili.
- Fare clic su Esegui. L'insieme di modelli viene collocato nello stream e nella palette Modelli nella parte superiore destra dello schermo.
- Fare doppio clic sull'insieme di modelli nello stream.

Esame dei dati

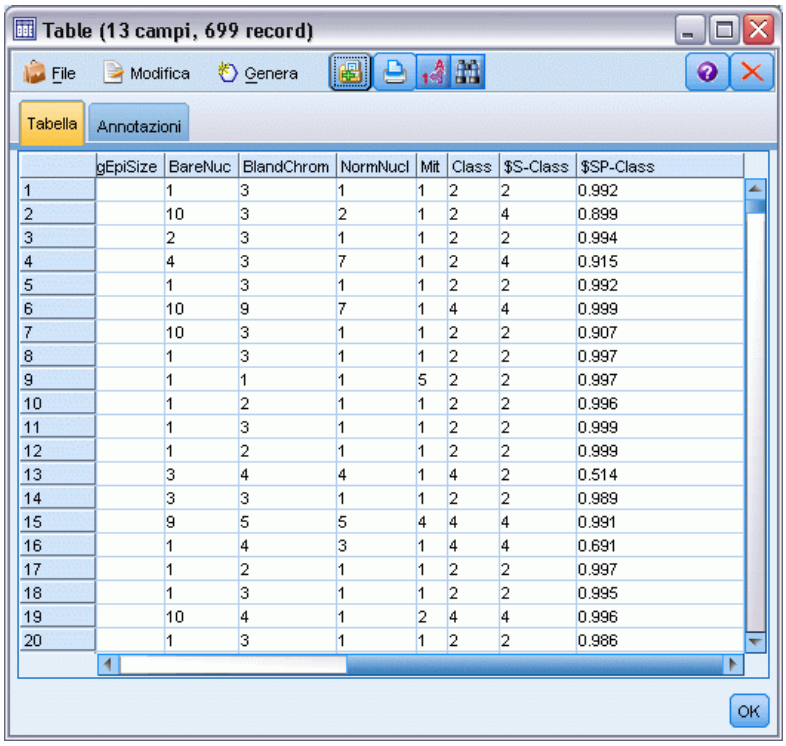
Figura 25-7
Grafico dell'importanza dei predittori



Nella scheda Modello, il grafico dell'importanza dei predittori mostra l'effetto relativo dei vari campi sulla previsione. Il grafico indica che *NucNudi* produce facilmente l'effetto massimo e anche *UnifForma* e *Agglutinazione* sono piuttosto significativi.

- Fare clic su OK.
- Collegare un nodo Tabella all'insieme di modelli *classe-rbf*.
- Aprire il nodo Tabella e fare clic su Esegui.

Figura 25-8
Campi aggiunti per i valori di previsione e confidenza



	gEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class	$S-Class	$SP-Class
1		1	3	1	1	2	2	0.992
2		10	3	2	1	2	4	0.899
3		2	3	1	1	2	2	0.994
4		4	3	7	1	2	4	0.915
5		1	3	1	1	2	2	0.992
6		10	9	7	1	4	4	0.999
7		10	3	1	1	2	2	0.907
8		1	3	1	1	2	2	0.997
9		1	1	1	5	2	2	0.997
10		1	2	1	1	2	2	0.996
11		1	3	1	1	2	2	0.999
12		1	2	1	1	2	2	0.999
13		3	4	4	1	4	2	0.514
14		3	3	1	1	2	2	0.989
15		9	5	5	4	4	4	0.991
16		1	4	3	1	4	4	0.691
17		1	2	1	1	2	2	0.997
18		1	3	1	1	2	2	0.995
19		10	4	1	2	4	4	0.996
20		1	3	1	1	2	2	0.986

▶ Il modello ha creato due campi aggiuntivi. Scorrere verso destra l'output della tabella per visualizzarli:

Nome nuovo campo	**Descrizione**
$S-Classe	Valore di *Classe* previsto dal modello.
$SP-Classe	Punteggio di propensione per questa previsione (poiché la verosimiglianza di questa previsione è vera, un valore da 0.0 a 1.0).

Osservando la tabella, si può notare che i punteggi di propensione (nella colonna *$SP-Classe*) per la maggior parte dei record sono ragionevolmente alti.

Tuttavia, vi sono alcune eccezioni significative; per esempio, il record del paziente 1041801 alla riga 13, dove il valore 0.514 è inaccettabilmente basso. Inoltre, confrontando il valore di *Classe* con quello di *$S-Classe*, è evidente che questo modello ha realizzato numerose previsioni errate, anche quando il punteggio di propensione era relativamente alto (per esempio, righe 2 e 4).

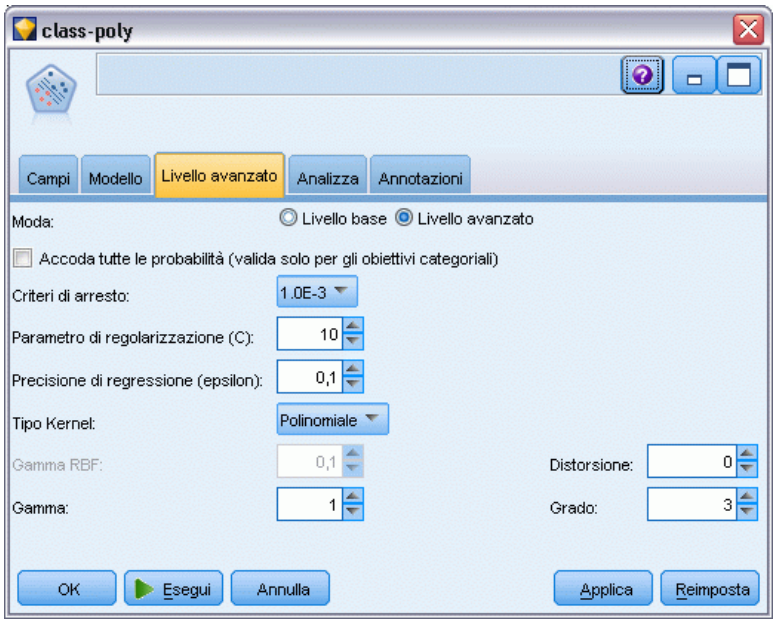
Si cercherà ora di migliorare i modelli scegliendo un altro tipo di funzione.

Tentativo con un'altra funzione

Figura 25-9
Impostazione di un nuovo nome per il modello

- Chiudere la finestra Output tabella.
- Collegare un secondo nodo Modelli SVM al nodo Tipo.
- Aprire il nuovo nodo SVM.
- Nella scheda Modello, scegliere Personalizzato e digitare *classe-poli* come nome del modello.

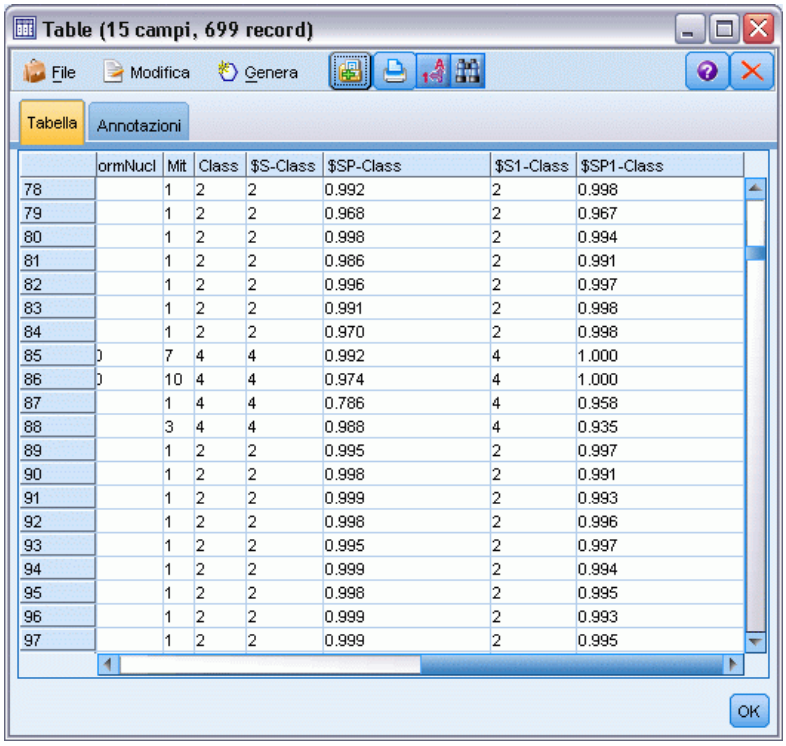
Figura 25-10
Impostazioni della scheda Livello avanzato per la modalità Polinomiale



- Nella scheda Livello avanzato, impostare la Modalità su Livello avanzato.
- Impostare il Tipo Kernel su Polinomiale e fare clic su Esegui. L'insieme di modelli *classe-poli* viene collocato nello stream e nella palette Modelli nella parte superiore destra dello schermo.
- Collegare l'insieme di modelli *classe-rbf* all'insieme di modelli *classe-poli* (scegliere Sostituisci nella finestra di avviso).
- Collegare un nodo Tabella all'insieme di modelli *classe-poli*.
- Aprire il nodo Tabella e fare clic su Esegui.

Confronto tra i risultati

Figura 25-11
Campi aggiunti per la funzione Polinomiale



	ormNucl	Mit	Class	$S-Class	$SP-Class	$S1-Class	$SP1-Class
78		1	2	2	0.992	2	0.998
79		1	2	2	0.968	2	0.967
80		1	2	2	0.998	2	0.994
81		1	2	2	0.986	2	0.991
82		1	2	2	0.996	2	0.997
83		1	2	2	0.991	2	0.998
84		1	2	2	0.970	2	0.998
85	)	7	4	4	0.992	4	1.000
86	)	10	4	4	0.974	4	1.000
87		1	4	4	0.786	4	0.958
88		3	4	4	0.988	4	0.935
89		1	2	2	0.995	2	0.997
90		1	2	2	0.998	2	0.991
91		1	2	2	0.999	2	0.993
92		1	2	2	0.998	2	0.996
93		1	2	2	0.995	2	0.997
94		1	2	2	0.999	2	0.994
95		1	2	2	0.998	2	0.995
96		1	2	2	0.999	2	0.993
97		1	2	2	0.999	2	0.995

- Scorrere verso destra l'output della tabella per visualizzare i campi aggiunti.

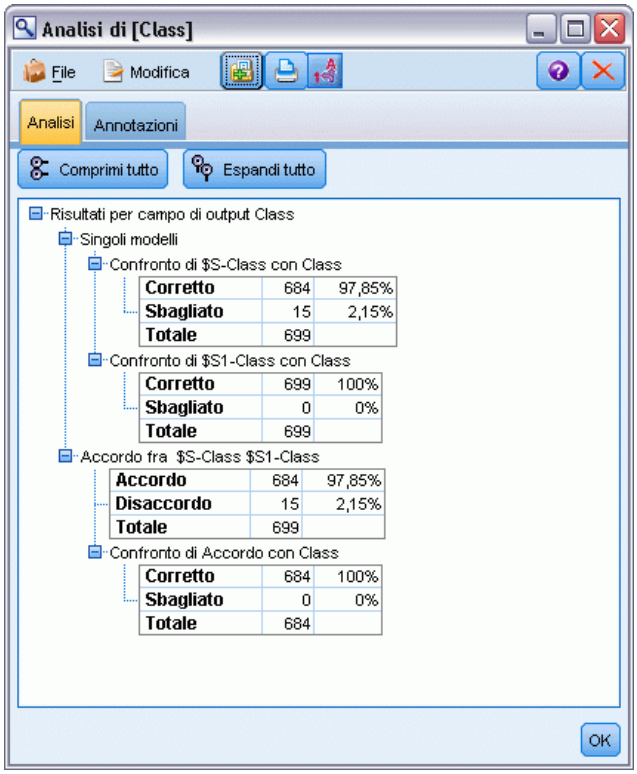
  I campi generati per il tipo di funzione Polinomiale sono denominati *$S1-Classe* e *$SP1-Classe*.

  I risultati di Polinomiale appaiono migliori. Molti punteggi di propensione hanno un valore pari o superiore a 0,995, un dato molto incoraggiante.

- Per confermare il miglioramento nel modello, collegare un nodo Analisi all'insieme di modelli *classe-poli*.

  Aprire il nodo Analisi e fare clic su Esegui.

Figura 25-12
nodo Analisi



Questa tecnica con il nodo Analisi consente di confrontare due o più insiemi di modelli dello stesso tipo. L'output del nodo Analisi mostra che la funzione RBF prevede correttamente il 97,85% dei casi, un dato piuttosto positivo. Tuttavia, l'output mostra che la funzione Polinomiale ha previsto correttamente la diagnosi in ogni singolo caso. In pratica, con un insieme di dati reale è poco probabile che si ottenga un grado di precisione pari al 100%, ma si potrà utilizzare il nodo Analisi per determinare se il modello è sufficientemente preciso per la propria applicazione.

Infatti, nessuno degli altri tipi di funzione (Sigmoide e Lineare) funziona bene come la funzione Polinomiale in questo particolare insieme di dati. Tuttavia, con un insieme di dati diverso, i risultati potrebbero essere facilmente diversi, perciò è sempre bene provare l'intera gamma di opzioni.

Riepilogo

Sono stati utilizzati diversi tipi di funzioni Kernel SVM per prevedere una classificazione a partire da numerosi attributi. Si è visto come Kernel diversi diano risultati diversi per lo stesso insieme di dati e come sia possibile misurare i miglioramenti di un modello rispetto all'altro.

Capitolo
26

Utilizzo della regressione di Cox per creare un modello del tempo di abbandono dei clienti

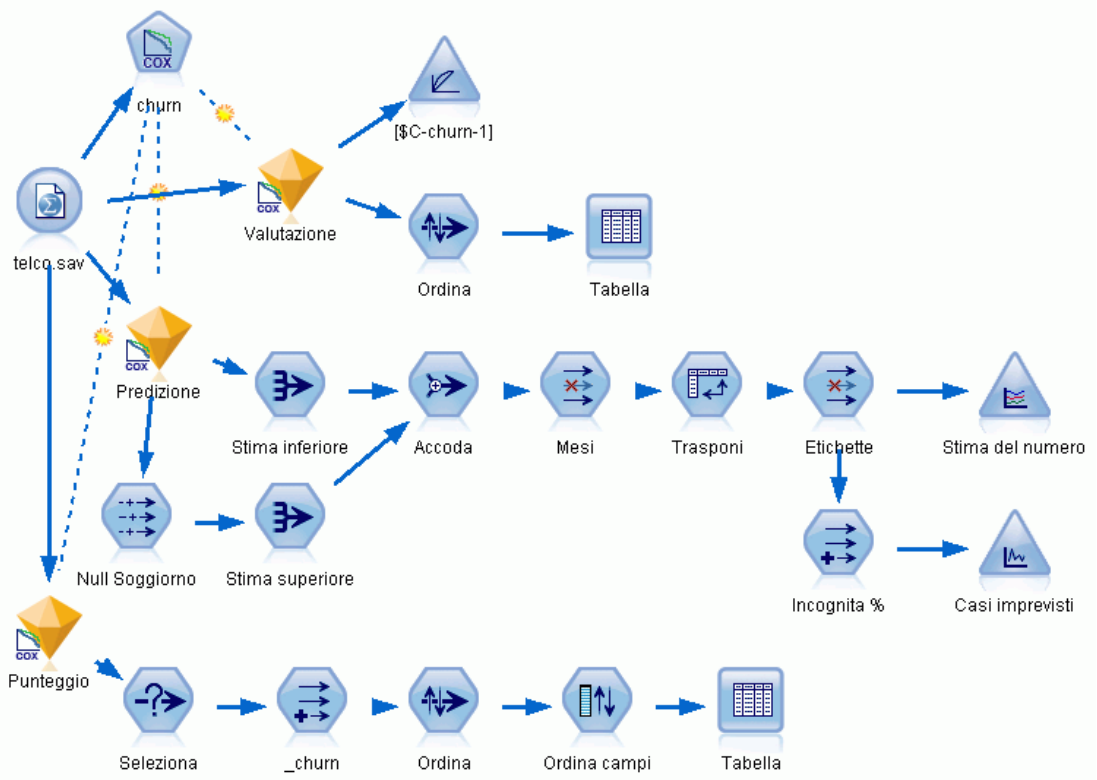
Nell'ambito degli sforzi per ridurre il tasso di abbandono dei clienti, una società di telecomunicazioni desidera creare un modello del "tempo di abbandono" per determinare i fattori associati ai clienti che passano rapidamente a un servizio concorrente. A tale scopo, viene selezionato un campione casuale di clienti e il relativo tempo trascorso in qualità di clienti (indipendentemente dal loro stato attuale di clienti) e svariati altri campi vengono estratti dal database.

In questo esempio viene utilizzato lo stream *telco_coxreg.str*, che fa riferimento al file di dati *telco.sav*. Il file di dati si trova nella cartella *Demos* e il file di stream nella sottocartella *streams*. Per ulteriori informazioni, vedere l'argomento Cartella Demos in il capitolo 1 in *Manuale dell'utente di IBM SPSS Modeler 15*.

Creazione di un modello idoneo

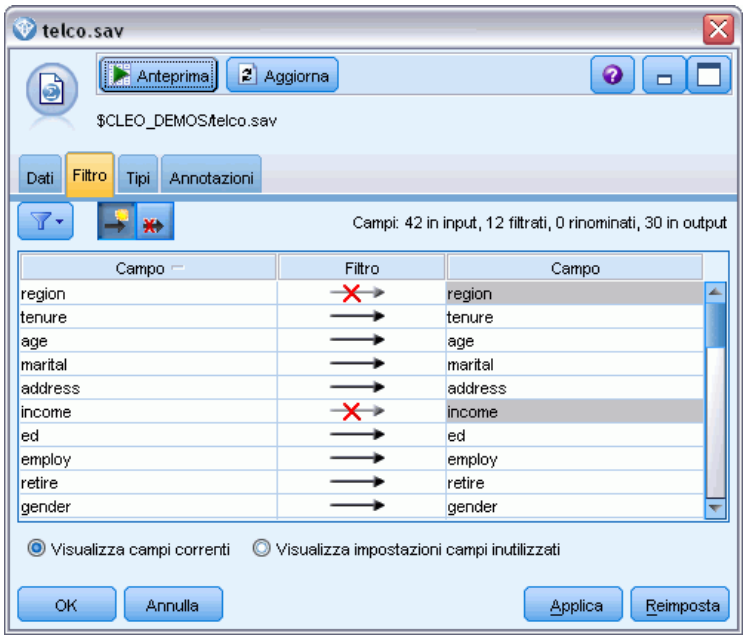
- Aggiungere un nodo di input File Statistics che punta a *telco.sav* nella cartella *Demos*.

Figura 26-1
Stream di esempio per analizzare il tempo di abbandono



▶ Nella scheda Filtro del nodo di input, escludere i campi *region*, *income*, i campi da *longten* a *wireten* e i campi da *loglong* a *logwire*.

Figura 26-2
Filtraggio dei campi superflui

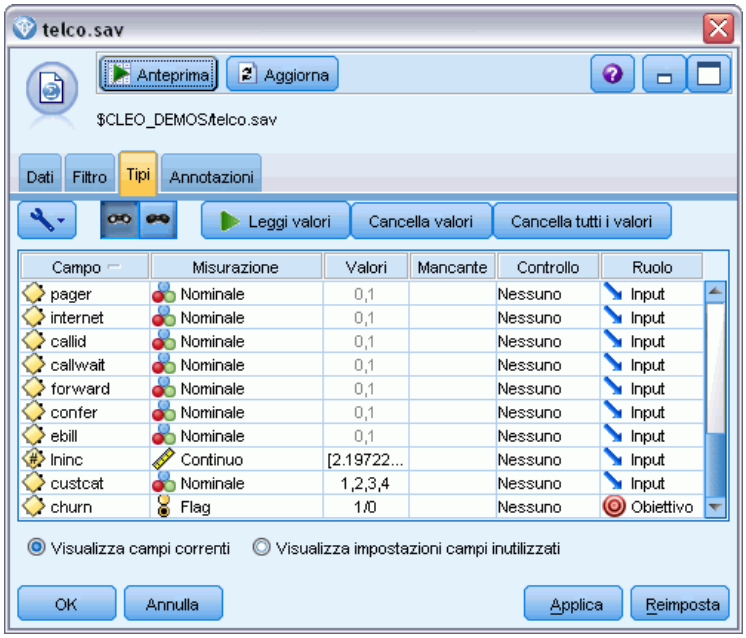


In alternativa, anziché escludere un campo, è possibile modificarne il ruolo in Nessuno nella scheda Tipi oppure selezionare i campi da utilizzare nel nodo Modelli.

▶ Nella scheda Tipi del nodo di input, impostare il ruolo del campo *churn* su Obiettivo e impostarne il livello di misurazione su Flag. Tutti gli altri campi dovrebbero avere il ruolo impostato su Input.

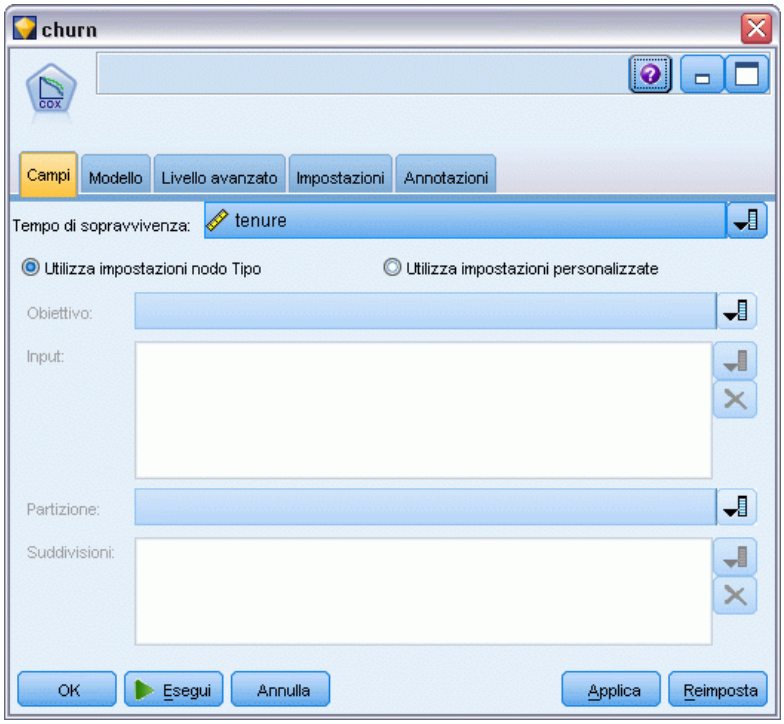
▶ Fare clic su Read Values (Leggi i valori) per creare un'istanza dei dati.

Figura 26-3
Impostazione del ruolo del campo



▶ Collegare un nodo Cox al nodo di input. Nella scheda Campi, selezionare *tenure* come variabile temporale di sopravvivenza.

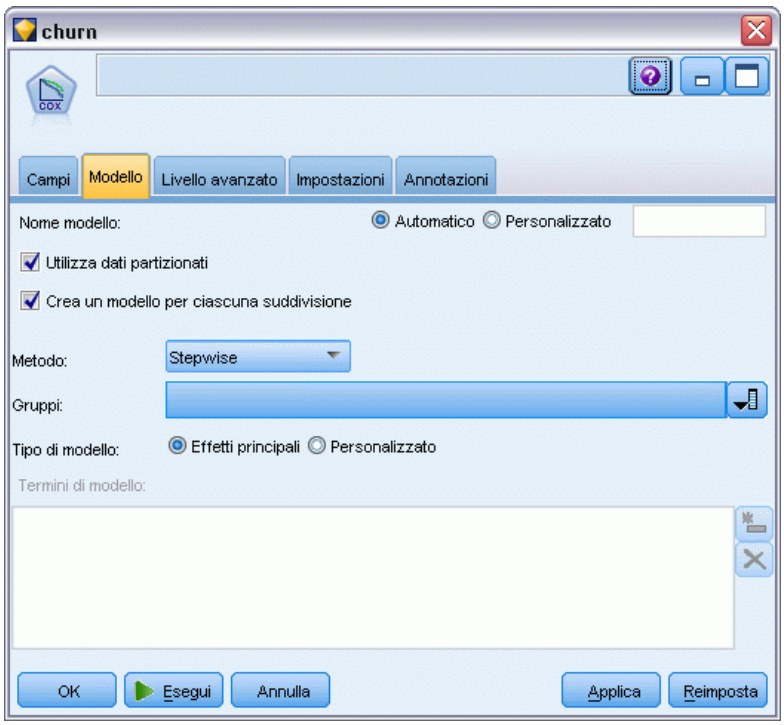
Figura 26-4
Scelta delle opzioni dei campi



▶ Fare clic sulla scheda Modello.

- Selezionare Stepwise come metodo di selezione delle variabili.

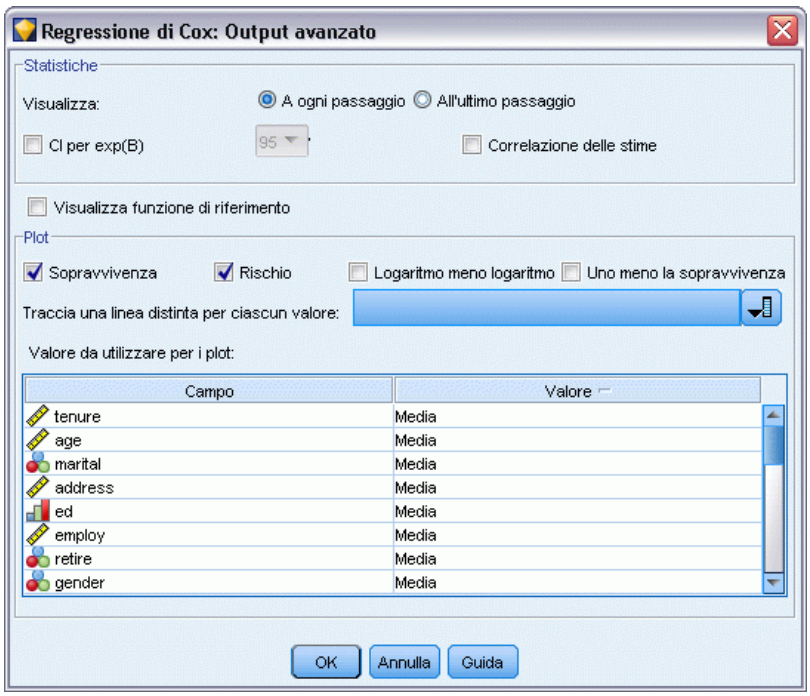
Figura 26-5
Scelta delle opzioni del modello



- Fare clic sulla scheda Livello avanzato e selezionare Livello avanzato per attivare le opzioni di modellazione avanzata.

▶ Fare clic su Output.

Figura 26-6
Scelta delle opzioni avanzate di output



▶ Selezionare Survival e Hazard come plot da generare, quindi fare clic su OK.

▶ Fare clic su Esegui per creare l'insieme di modelli, che viene aggiunto allo stream e alla palette Modelli nell'angolo superiore destro. Per visualizzare i relativi dettagli, fare doppio clic sull'insieme di modelli nello stream. Analizzare innanzitutto la scheda Output avanzato.

Casi censurati

Figura 26-7
Riepilogo dell'elaborazione dei casi

		N	Percentuale
Casi disponibili nell'analisi	**Evento(a)**	274	27,4%
	Troncati	726	72,6%
	Totale	1000	100,0%
Casi esclusi	**Casi con valori mancanti**	0	,0%
	Casi con tempo negativo	0	,0%
	Casi troncati prima del primo evento in uno strato	0	,0%
	Totale	0	,0%
Totale		1000	100,0%

a. Variabile dipendente: Months with service

La variabile di stato identifica se l'evento si è verificato per un determinato caso. Se l'evento non si è verificato, il caso si dice censurato. I casi censurati non vengono utilizzati nel calcolo dei coefficienti di regressione, ma sono utilizzati per calcolare il rischio di riferimento. Il riepilogo dell'elaborazione del caso mostra che sono stati censurati 726 casi. Si tratta dei clienti che non sono passati a un altro operatore.

Codifica delle variabili categoriali

Figura 26-8
Codifica delle variabili categoriali

		Frequenza	(1)(s)	(2)	(3)	(4)
marital(t)	0=Unmarried	505	1			
	1=Married	495	0			
ed(t)	1=Did not complete high school	204	1	0	0	0
	2=High school degree	287	0	1	0	0
	3=Some college	209	0	0	1	0
	4=College degree	234	0	0	0	1
	5=Post-undergraduate degree	66	0	0	0	0
retire(t)	0=No	953	1			
	1=Yes	47	0			
gender(t)	0=Male	483	1			
	1=Female	517	0			
tollfree(t)	0=No	526	1			
	1=Yes	474	0			
equip(t)	0=No	614	1			
	1=Yes	386	0			
callcard(t)	0=No	322	1			
	1=Yes	678	0			
wireless(t)	0=No	704	1			
	1=Yes	296	0			
multline(t)	0=No	525	1			
	1=Yes	475	0			
voice(t)	0=No	696	1			
	1=Yes	304	0			
pager(t)	0=No	739	1			
	1=Yes	261	0			
internet(t)	0=No	632	1			
	1=Yes	368	0			
callid(t)	0=No	519	1			
	1=Yes	481	0			
callwait(t)	0=No	515	1			
	1=Yes	485	0			
forward(t)	0=No	507	1			
	1=Yes	493	0			
confer(t)	0=No	498	1			
	1=Yes	502	0			
ebill(t)	0=No	629	1			
	1=Yes	371	0			
custcat(t)	1=Basic service	266	1	0	0	
	2=E-service	217	0	1	0	
	3=Plus service	281	0	0	1	
	4=Total service	236	0	0	0	

La codifica delle variabili categoriali è utile per interpretare i coefficienti di regressione della covariate categoriali, in particolare delle variabili dicotomiche. Per default, la categoria di riferimento è l'"ultima" categoria. Ciò significa, per esempio, che se anche i clienti *Married* hanno un valore di variabile pari a 1 nel file di dati, vengono codificati come 0 ai fini della regressione.

Selezione delle variabili

Figura 26-9
Test omnibus

Passo	-2 Log verosimiglianza	Globale (punteggio)			Cambiamento dal passo precedente			Cambiamento dal blocco precedente		
		Chi quadrato	df	Sig.	Chi quadrato	df	Sig.	Chi quadrato	df	Sig.
1(c)	3392,536	162,303	1	,000	133,828	1	,000	133,828	1	,000
2(d)	3087,314	249,392	2	,000	305,222	1	,000	439,050	2	,000
3(e)	3027,085	328,426	3	,000	60,229	1	,000	499,279	3	,000
4(f)	2990,790	347,197	4	,000	36,294	1	,000	535,574	4	,000
5(g)	2973,790	362,673	5	,000	17,000	1	,000	552,574	5	,000
6(h)	2958,796	376,140	6	,000	14,994	1	,000	567,568	6	,000
7(i)	2945,503	384,717	7	,000	13,293	1	,000	580,861	7	,000
8(j)	2936,993	417,341	8	,000	8,510	1	,004	589,371	8	,000
9(k)	2926,000	423,911	9	,000	10,994	1	,001	600,364	9	,000
10(l)	2917,551	428,078	10	,000	8,449	1	,004	608,813	10	,000
11(m)	2913,308	436,837	11	,000	4,243	1	,039	613,056	11	,000
12(n)	2908,078	440,158	12	,000	5,230	1	,022	618,286	12	,000

a. Blocco iniziale numero 0, funzione log verosimiglianza iniziale: -2 Log verosimiglianza: 3526,364

b. Blocco iniziale numero 1. Metodo = Stepwise in avanti (Rapporto di verosimiglianza)

c. Variabili specificate al passo 1: callcard

d. Variabili specificate al passo 2: longmon

e. Variabili specificate al passo 3: equip

f. Variabili specificate al passo 4: employ

g. Variabili specificate al passo 5: multline

h. Variabili specificate al passo 6: voice

i. Variabili specificate al passo 7: address

j. Variabili specificate al passo 8: equipmon

k. Variabili specificate al passo 9: ebill

l. Variabili specificate al passo 10: callid

m. Variabili specificate al passo 11: internet

n. Variabili specificate al passo 12: reside

Il processo di generazione del modello utilizza un algoritmo stepwise in avanti. I test omnibus misurano il livello delle prestazioni del modello. La variazione chi-quadrato dal passaggio precedente è la differenza tra la log-verosimiglianza −2 del modello nel passaggio precedente e il passaggio corrente. Se il passaggio riguardava l'aggiunta di una variabile, l'inclusione ha senso se la significatività della variazione è inferiore a 0.05. Se il passaggio riguardava la rimozione di una variabile, l'esclusione ha senso se la significatività della variazione è superiore a 0.10. In dodici passaggi vengono aggiunte al modello dodici variabili.

Figura 26-10
Variabili incluse nell'equazione (solo passaggio 12)

		B	SE	Wald	df	Sig.	Exp(B)
Passo 12	address	-,035	,009	14,543	1	,000	,966
	employ	-,051	,010	25,767	1	,000	,950
	reside	-,103	,046	5,037	1	,025	,902
	equip	-1,948	,381	26,180	1	,000	,143
	callcard	,777	,151	26,451	1	,000	2,175
	longmon	-,233	,022	115,619	1	,000	,792
	equipmon	-,042	,011	15,377	1	,000	,959
	multline	,612	,145	17,854	1	,000	1,844
	voice	-,501	,157	10,197	1	,001	,606
	internet	-,362	,160	5,114	1	,024	,697
	callid	-,464	,148	9,790	1	,002	,629
	ebill	-,399	,156	6,557	1	,010	,671

Il modello finale include *address*, *employ*, *reside*, *equip*, *callcard*, *longmon*, *equipmon*, *multline*, *voice*, *internet*, *callid* e*ebill*. Per comprendere gli effetti dei singoli predittori, esaminare Exp(B), che può essere interpretato come la variazione prevista nel rischio per un aumento di un'unità nel predittore.

- Il valore di Exp(B) per *address* significa che il rischio di abbandono è ridotto del 100%−(100%×0.966)=3.4% per ogni anno in cui un cliente ha vissuto allo stesso indirizzo. Il rischio di abbandono per un cliente che ha vissuto allo stesso indirizzo per cinque anni è ridotto del $100\%-(100\%\times 0.966^5)=15.88\%$.
- Il valore di Exp(B) per *callcard* significa che il rischio di abbandono per un cliente che non sottoscrive il servizio carta telefonica è 2.175 volte superiore a quello di un cliente che utilizza tale servizio. Tenere presente che in base alla codifica delle variabili categoriali *No* = 1 per la regressione.
- Il valore di Exp(B) per *internet* significa che il rischio di abbandono per un cliente che non sottoscrive il servizio Internet è 0.697 volte superiore a quello di un cliente che utilizza tale servizio. Questo dato è alquanto preoccupante perché significa che i clienti che utilizzano il servizio Internet passano a un altro operatore più rapidamente di quelli che non lo utilizzano.

Figura 26-11
Variabili escluse dal modello (solo passaggio 12)

Passo 12	age	,122	1	,726
	marital	,648	1	,421
	ed	6,328	4	,176
	ed(1)	,007	1	,934
	ed(2)	,203	1	,652
	ed(3)	,835	1	,361
	ed(4)	5,773	1	,016
	retire	,013	1	,908
	gender	,214	1	,644
	tollfree	3,243	1	,072
	wireless	,668	1	,414
	tollmon	,000	1	,987
	cardmon	3,163	1	,075
	wiremon	1,084	1	,298
	pager	1,808	1	,179
	callwait	,266	1	,606
	forward	2,201	1	,138
	confer	2,568	1	,109
	lninc	2,853	1	,091
	custcat	,864	3	,834
	custcat(1)	,466	1	,495
	custcat(2)	,450	1	,502
	custcat(3)	,019	1	,889

Le variabili escluse dal modello hanno tutte statistiche di punteggio con valori di significatività superiori a 0.05. Tuttavia, i valori di significatività per *tollfree* e *cardmon*, sebbene non inferiori a 0.05, vi si avvicinano molto. Potrebbe essere interessante approfondirli con ulteriori studi.

Medie delle covariate

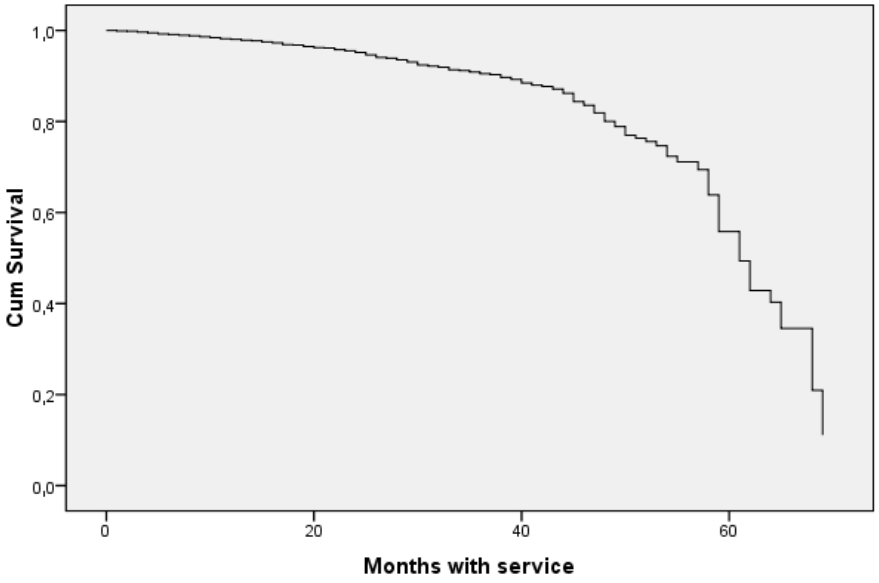
Figura 26-12
Medie delle covariate

	Media
age	41,684
marital	,505
address	11,551
ed(1)	,204
ed(2)	,287
ed(3)	,209
ed(4)	,234
employ	10,987
retire	,953
gender	,483
reside	2,331
tollfree	,526
equip	,614
callcard	,322
wireless	,704
longmon	11,723
tollmon	13,274
equipmon	14,220
cardmon	13,781
wiremon	11,584
multline	,525
voice	,696
pager	,739
internet	,632
callid	,519
callwait	,515
forward	,507
confer	,498
ebill	,629
lninc	3,957
custcat(1)	,266
custcat(2)	,217
custcat(3)	,281

In questa tabella è riportato il valore medio di ogni variabile predittore. Questa tabella è un riferimento utile quando si esaminano i grafici plot di sopravvivenza, che sono costruiti per i valori medi. Si noti tuttavia che il cliente "medio" non esiste effettivamente quando si esaminano le medie delle variabili indicatore per i predittori categoriali. Anche con tutti i predittori di scala, è improbabile trovare un cliente i cui valori delle covariate siano tutti vicini alla media. Se si desidera visualizzare la curva di sopravvivenza per un determinato caso, è possibile cambiare i valori delle covariate ai quali la curva di sopravvivenza è rappresentata graficamente nella finestra di dialogo Plot. Se si desidera visualizzare la curva di sopravvivenza per un determinato caso, è possibile cambiare i valori delle covariate ai quali la curva di sopravvivenza è rappresentata graficamente nel gruppo Plot della finestra di dialogo Output avanzato.

Curva di sopravvivenza

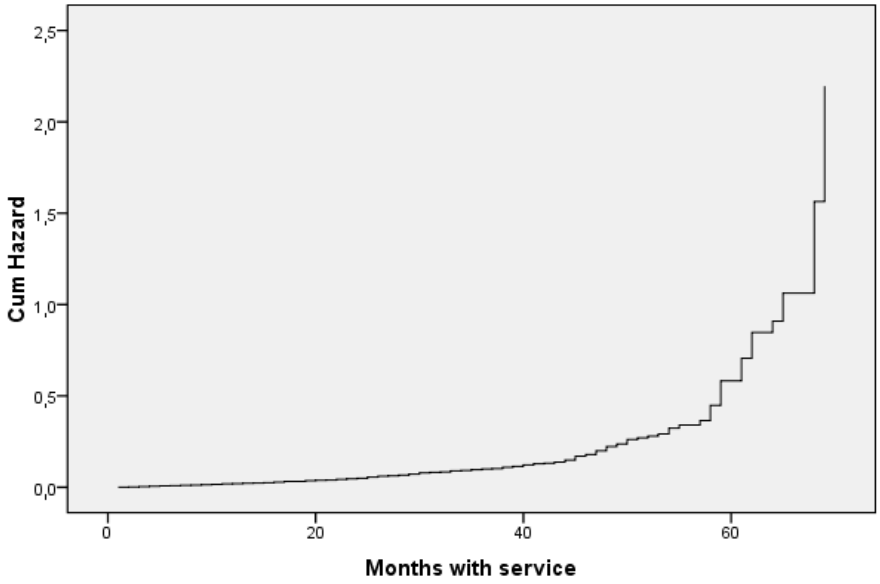
Figura 26-13
Curva di sopravvivenza per cliente "medio"



La curva di sopravvivenza di base è una visualizzazione grafica del tempo di abbandono previsto dal modello per il cliente "medio". Sull'asse orizzontale è indicato il tempo all'evento. Sull'asse verticale è indicata la probabilità di sopravvivenza. Pertanto, qualsiasi punto sulla curva di sopravvivenza indica la probabilità che il cliente "medio" rimanga cliente trascorso tale tempo. Trascorsi 55 mesi, la curva di sopravvivenza diventa meno uniforme. I clienti rimasti fedeli alla società così a lungo sono di meno, pertanto le informazioni disponibili sono inferiori e la curva risulta a blocchi.

Curva di rischio

Figura 26-14
Curva di rischio per cliente "medio"

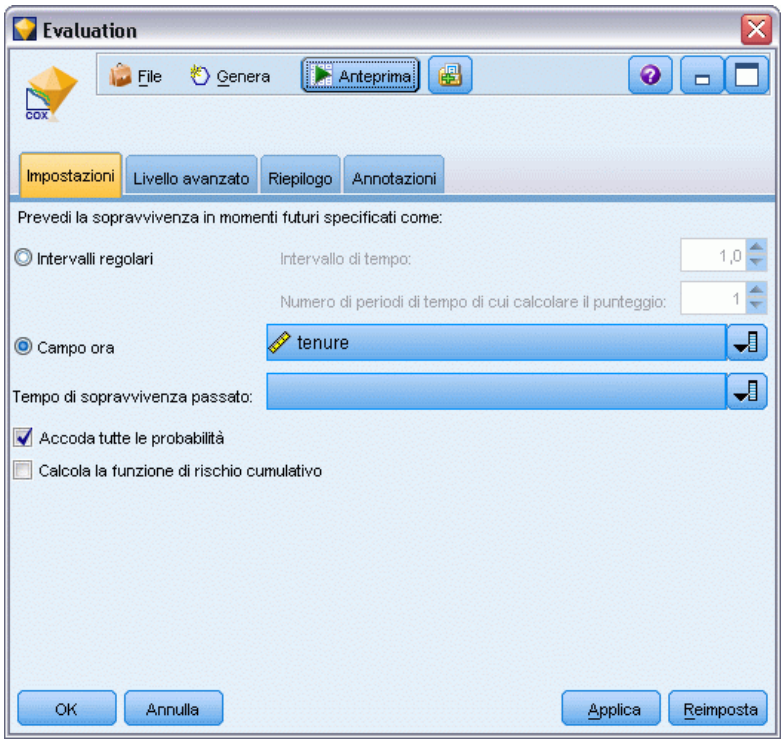


La curva di rischio di base è una visualizzazione grafica del potenziale di abbandono cumulato previsto dal modello per il cliente "medio". Sull'asse orizzontale è indicato il tempo all'evento. Sull'asse verticale è indicato il rischio cumulato, pari al logaritmo negativo della probabilità di sopravvivenza. Trascorsi 55 mesi, la curva di rischio, analogamente alla curva di sopravvivenza, diventa meno omogenea, per lo stesso motivo.

Valutazione

I metodi di selezione stepwise garantiscono che il modello avrà solo predittori "statisticamente significativi", ma non assicura che il modello sia effettivamente efficace nel prevedere l'obiettivo. A questo scopo è necessario analizzare i record di cui è stato calcolato il punteggio.

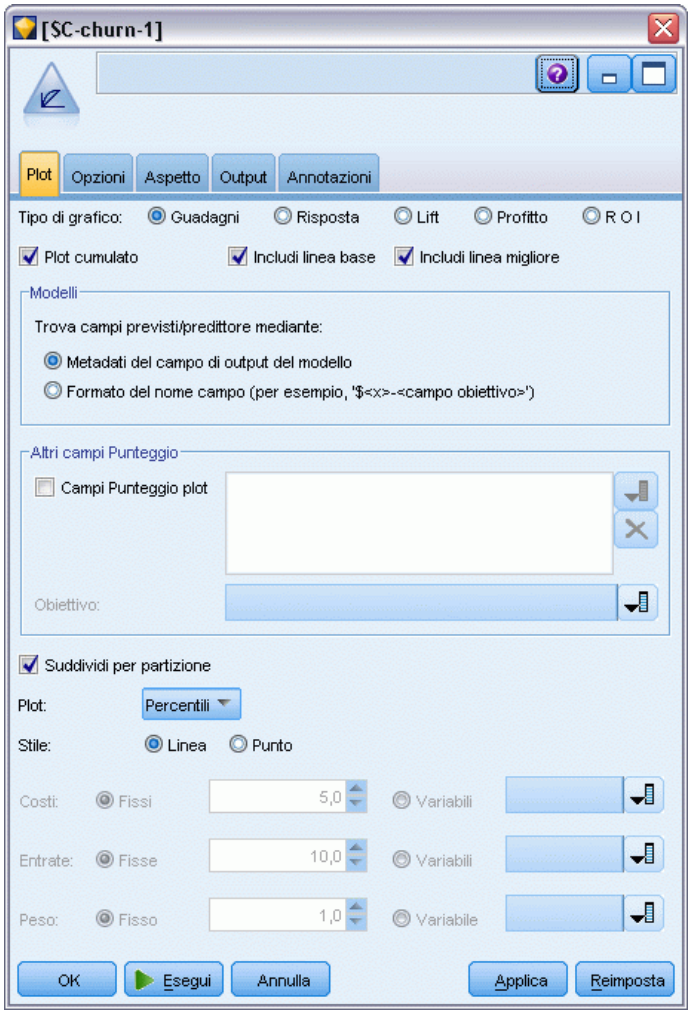
Figura 26-15
Insieme di modelli Regressione di Cox: scheda Impostazioni



- Posizionare l'insieme di modelli nell'area di disegno e collegarlo al nodo di input. Aprire quindi l'insieme di modelli e fare clic sulla scheda Impostazioni.
- Selezionare Campo ora e specificare *tenure*. Di ogni record verrà calcolato il punteggio rispetto alla rispettiva durata (tenure).
- Selezionare Accoda tutte le probabilità.

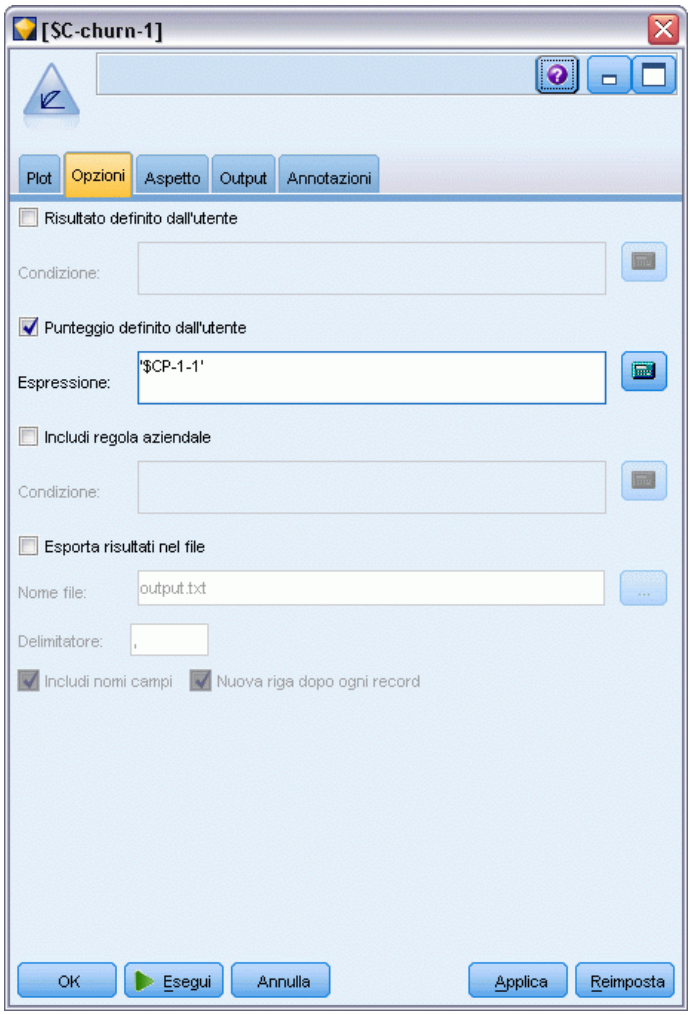
In questo modo, nel calcolo dei punteggi viene utilizzato 0.5 come valore di interruzione; ciò significa che se la propensione all'abbandono dei clienti è superiore a 0.5, a questi clienti viene attribuito il punteggio di "cliente che cambia operatore". Non vi è nulla di speciale in questo numero e un valore di interruzione diverso potrebbe fornire risultati migliori. Per scegliere un valore di interruzione adeguato, è possibile utilizzare un nodo Valutazione.

Figura 26-16
Nodo Valutazione: scheda Plot



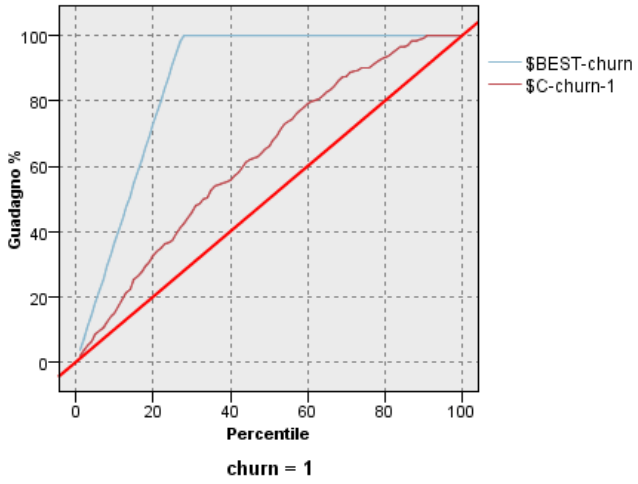
- Collegare un nodo Valutazione all'insieme di modelli. Nella scheda Plot, selezionare Includi linea migliore.
- Fare clic sulla scheda Opzioni.

Figura 26-17
Nodo Valutazione: Scheda Opzioni



- Selezionare Punteggio definito dall'utente e digitare '$CP-1-1' come espressione. Questo è un campo generato dal modello che corrisponde alla propensione all'abbandono.
- Fare clic su Esegui.

Figura 26-18
Grafico dei guadagni



Il grafico dei guadagni cumulativi mostra la percentuale del numero totale dei casi in una determinata categoria ottenuta definendo come target una percentuale del numero totale dei casi. Per esempio, un punto sulla curva è a (10%, 15%). Ciò significa che se si calcola il punteggio di un insieme di dati con il modello e si ordinano tutti i casi in base alla propensione all'abbandono prevista, ci si attende che il 10% superiore contenga approssimativamente il 15% di tutti i casi che costituiscono la categoria *1* (clienti che cambiano operatore). Analogamente, il 60% superiore contiene approssimativamente il 79.2% dei clienti che cambiano operatore. Se si seleziona il 100% dell'insieme di dati di cui si calcola il punteggio, si ottengono tutti i clienti che cambiano operatore nell'insieme di dati.

La linea diagonale è la curva "di riferimento". Selezionando a caso il 20% dei record dall'insieme di dati di cui si calcola il punteggio, ci si attenderà di "guadagnare" approssimativamente il 20% di tutti i record che costituiscono la categoria *1*. Più in alto si trova la linea di base in cui si trova una curva, maggiore è il guadagno. La linea "migliore" mostra la curva di un modello "perfetto", che assegna un punteggio superiore di propensione all'abbandono a ogni cliente che cambia operatore rispetto a quelli che non lo cambiano. È possibile utilizzare il grafico dei guadagni cumulativi per consentire di scegliere un riferimento della classificazione scegliendo una percentuale che corrisponda a un guadagno preferibile, quindi associando tale percentuale al valore di riferimento appropriato.

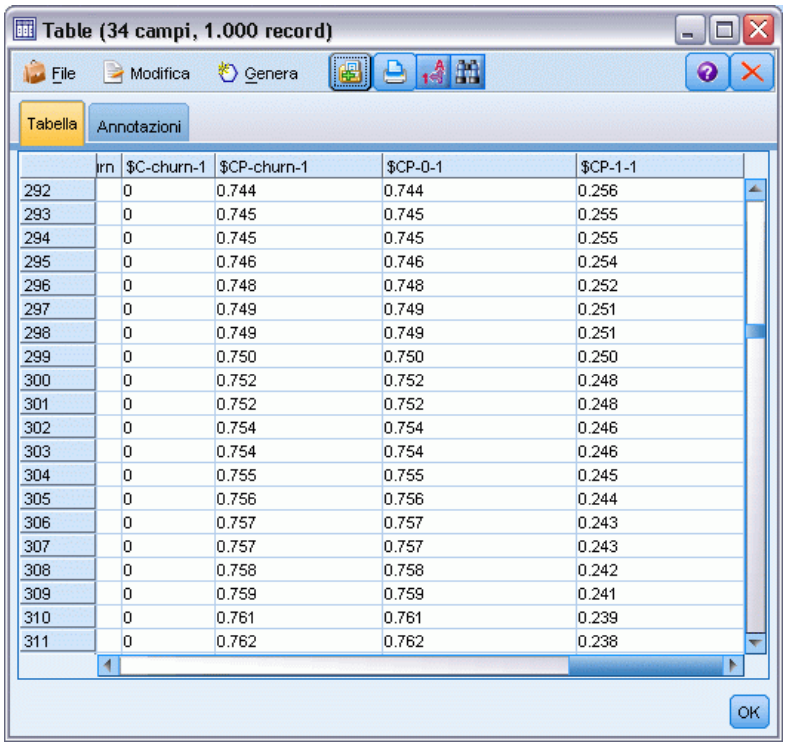
La definizione di guadagno "preferibile" dipende dal costo degli errori di Tipo I e di Tipo II. In altre parole, qual è il costo che deriva dal classificare un cliente che cambia operatore come cliente che non cambia operatore (Tipo I)? E ancora, qual è il costo che deriva dal classificare un cliente che non cambia operatore come cliente che cambia operatore (Tipo II)? Se la preoccupazione principale è la fidelizzazione dei clienti, si vorrà diminuire l'errore di Tipo I. Nel grafico dei guadagni cumulato, ciò potrebbe corrispondere a un migliore servizio di assistenza ai clienti nel 60% superiore della propensione prevista di *1*, che cattura il 79.2% dei possibili clienti che cambiano operatore, ma questa strategia è dispendiosa in termini di tempo e risorse, che potrebbero essere invece utilizzati per acquisire nuovi clienti. Se la priorità è invece ridurre il costo del mantenimento della base clienti attuale, si vorrà ridurre l'errore di Tipo II. Nel grafico, ciò potrebbe corrispondere a un migliore servizio di assistenza ai clienti nel 20% superiore, che

cattura il 32.5% dei possibili clienti che cambiano operatore. Generalmente, entrambi gli obiettivi sono importanti e pertanto si dovrà scegliere una regola per la classificazione dei clienti che offra il miglior compromesso tra sensibilità e specificità.

Figura 26-19
Nodo Ordina: scheda Impostazioni

- Si supponga che il guadagno desiderabile sia stato fissato al 45.6%, che corrisponde al 30% superiore dei record. Per trovare il valore di interruzione appropriato per la classificazione, collegare un nodo Ordina all'insieme di modelli.
- Nella scheda Impostazioni, scegliere di ordinare per *$CP-1-1* in ordine decrescente e fare clic su OK.

Figura 26-20
Tabella



	irn	$C-churn-1	$CP-churn-1	$CP-0-1	$CP-1-1
292		0	0.744	0.744	0.256
293		0	0.745	0.745	0.255
294		0	0.745	0.745	0.255
295		0	0.746	0.746	0.254
296		0	0.748	0.748	0.252
297		0	0.749	0.749	0.251
298		0	0.749	0.749	0.251
299		0	0.750	0.750	0.250
300		0	0.752	0.752	0.248
301		0	0.752	0.752	0.248
302		0	0.754	0.754	0.246
303		0	0.754	0.754	0.246
304		0	0.755	0.755	0.245
305		0	0.756	0.756	0.244
306		0	0.757	0.757	0.243
307		0	0.757	0.757	0.243
308		0	0.758	0.758	0.242
309		0	0.759	0.759	0.241
310		0	0.761	0.761	0.239
311		0	0.762	0.762	0.238

- Collegare un nodo Tabella al nodo Ordina.
- Aprire il nodo Tabella e fare clic su Esegui.

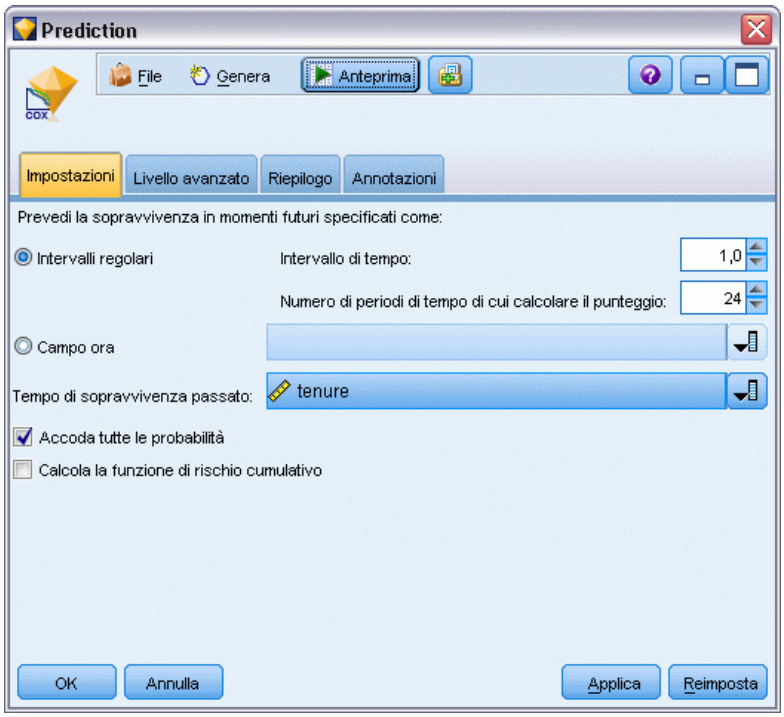
Scorrendo i risultati verso il basso si noterà che il valore *$CP-1-1* è 0,248 per il 300esimo record. Se si utilizza 0.248 come valore di interruzione per la classificazione, si ottiene approssimativamente il 30% dei clienti calcolati come clienti che cambiano operatore, catturando così circa il 45% di tutti i clienti effettivi che cambiano operatore.

Verifica del numero previsto di clienti mantenuti

Una volta che si è soddisfatti del modello, si vorrà verificare il numero previsto di clienti dell'insieme di dati che vengono mantenuti nel corso dei due anni successivi. I valori nulli, che sono i clienti la cui durata totale (tempo futuro + *tenure*) non è compresa nell'intervallo dei tempi di sopravvivenza nei dati utilizzati per addestrare il modello, costituiscono una sfida interessante. Un metodo per gestirli consiste nel creare due insiemi di previsioni: una in cui si presume che i valori nulli abbiano lasciato la società e un'altra in cui si presume che gli stessi le siano rimasti

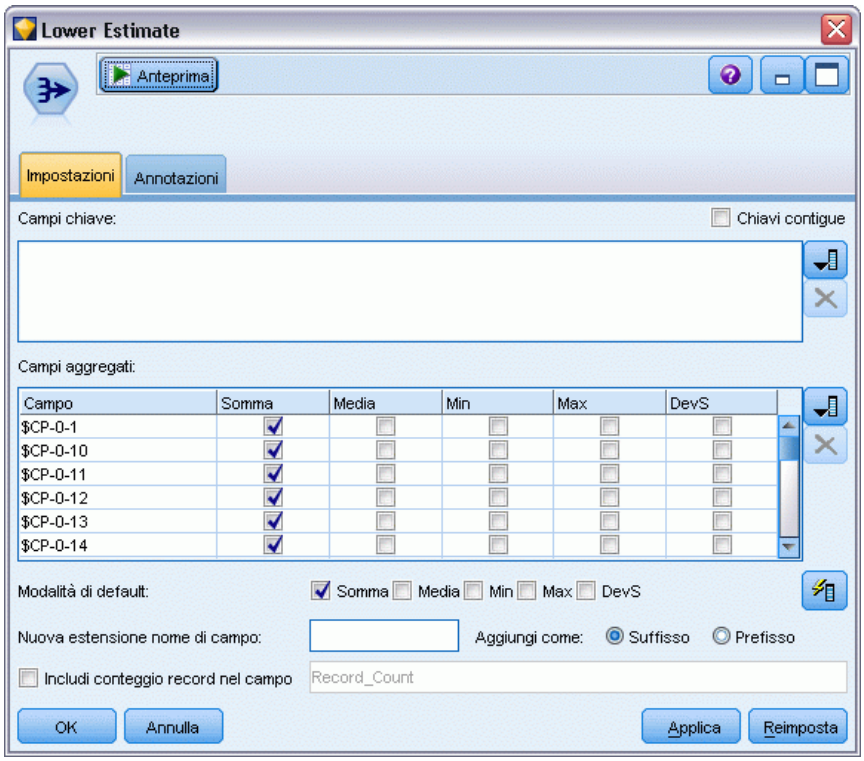
fedeli. In questo modo è possibile stabilire un limite superiore e inferiore del numero previsto di clienti mantenuti.

Figura 26-21
Insieme di modelli Regressione di Cox: scheda Impostazioni



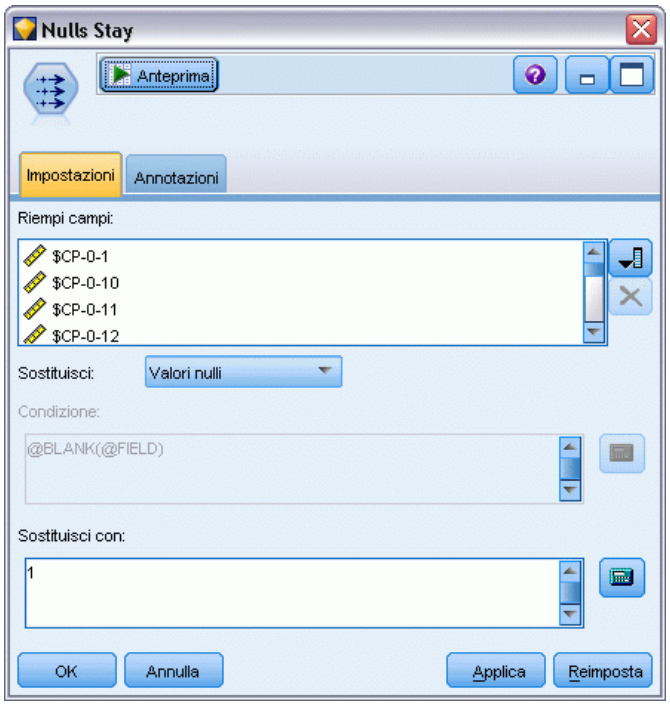
- Fare doppio clic sull'insieme di modelli nella palette Modelli (o copiare e incollare l'insieme nell'area di disegno dello stream) e collegare il nuovo insieme al nodo Input.
- Aprire l'insieme di modelli sulla scheda Impostazioni.
- Assicurarsi che sia selezionata l'opzione Intervalli regolari e specificare 1,0 come intervallo temporale e 24 come numero di periodi di cui calcolare il punteggio. In questo modo, di ogni record verrà calcolato il punteggio per ognuno dei 24 mesi successivi.
- Selezionare *tenure* come campo per specificare il tempo di sopravvivenza passato. L'algoritmo di calcolo del punteggio terrà conto della durata di ogni cliente come cliente della società.
- Selezionare Accoda tutte le probabilità.

Figura 26-22
Nodo Aggregazione: scheda Impostazioni



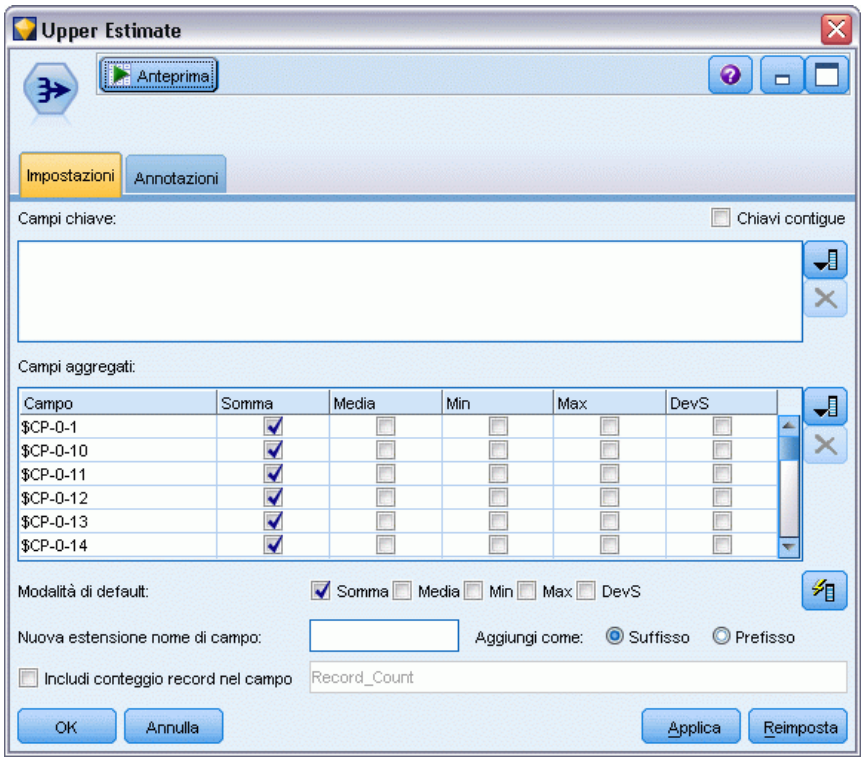
- Collegare un nodo Aggregazione all'insieme di modelli. Nella scheda Impostazioni, deselezionare Media come modalità di default.
- Come campi da aggregare selezionare da *$CP-0-1* a *$CP-0-24*, i campi con formato *$CP-0-n*. Questa operazione risulta più semplice se nella finestra di dialogo Seleziona campi i campi vengono ordinati per Nome, ovvero in ordine alfabetico.
- Deselezionare Includi conteggio record nel campo.
- Fare clic su OK. Questo nodo crea le previsioni "limite inferiore".

Figura 26-23
Nodo Riempimento: scheda Impostazioni



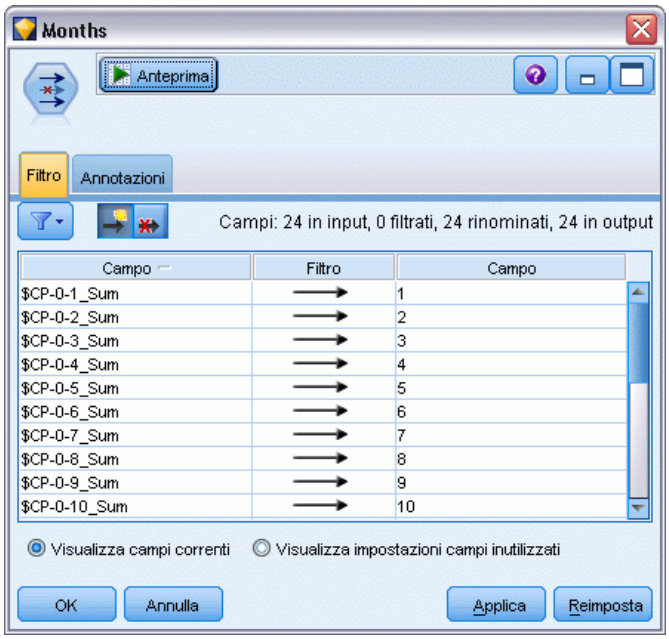
- Collegare un nodo Riempimento all'insieme di modelli Regressione di Cox a cui è stato appena collegato il nodo Aggregazione. Nella scheda Impostazioni, selezionare come campi da riempire da *$CP-0-1* a *$CP-0-24*, i campi con formato *$CP-0-n*. Questa operazione risulta più semplice se nella finestra di dialogo Seleziona campi i campi vengono ordinati per Nome, ovvero in ordine alfabetico.
- Scegliere di sostituire i Valori nulli con il valore 1.
- Fare clic su OK.

Figura 26-24
Nodo Aggregazione: scheda Impostazioni



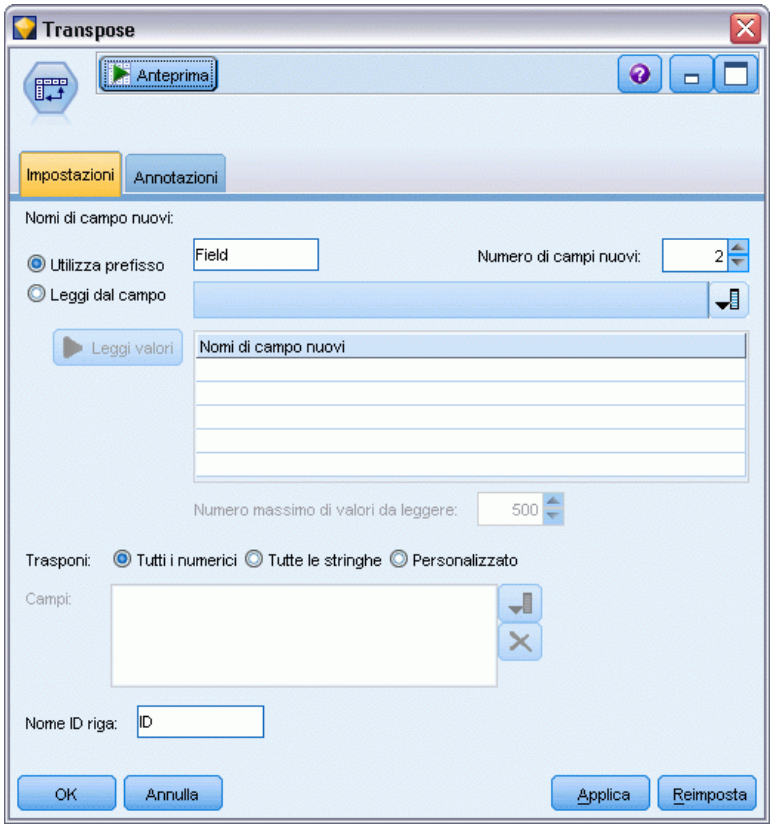
- Collegare un nodo Aggregazione al nodo Riempimento. Nella scheda Impostazioni, deselezionare Media come modalità di default.
- Come campi da aggregare selezionare da *$CP-0-1* a *$CP-0-24*, i campi con formato *$CP-0-n*. Questa operazione risulta più semplice se nella finestra di dialogo Seleziona campi i campi vengono ordinati per Nome, ovvero in ordine alfabetico.
- Deselezionare Includi conteggio record nel campo.
- Fare clic su OK. Questo nodo crea le previsioni “limite superiore”.

Figura 26-25
Nodo Filtro: scheda Impostazioni



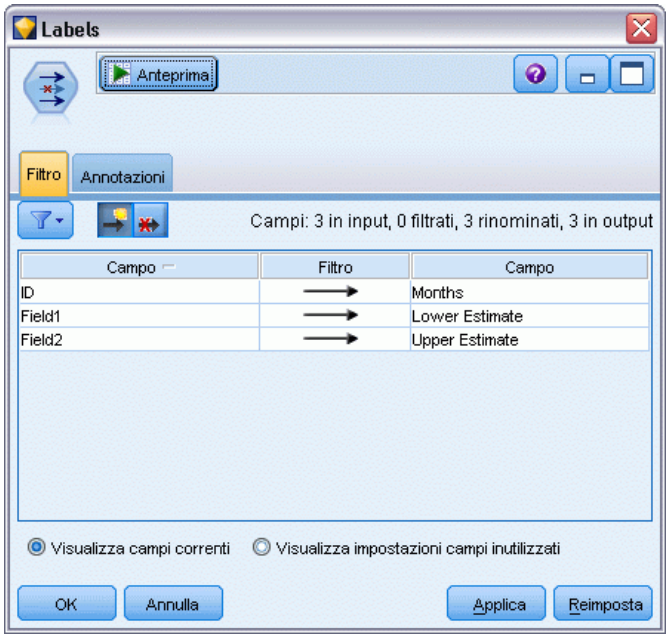
- Collegare un nodo Accoda ai due nodi Aggregazione, quindi collegare un nodo Filtro al nodo Accoda.
- Nella scheda Impostazioni del nodo Filtro, ridenominare i campi da *1* a *24*. Tramite l'utilizzo di un nodo Trasponi, questi nomi di campo diventeranno i valori dell'asse *x* nei grafici a valle.

Figura 26-26
Nodo Trasponi: scheda Impostazioni



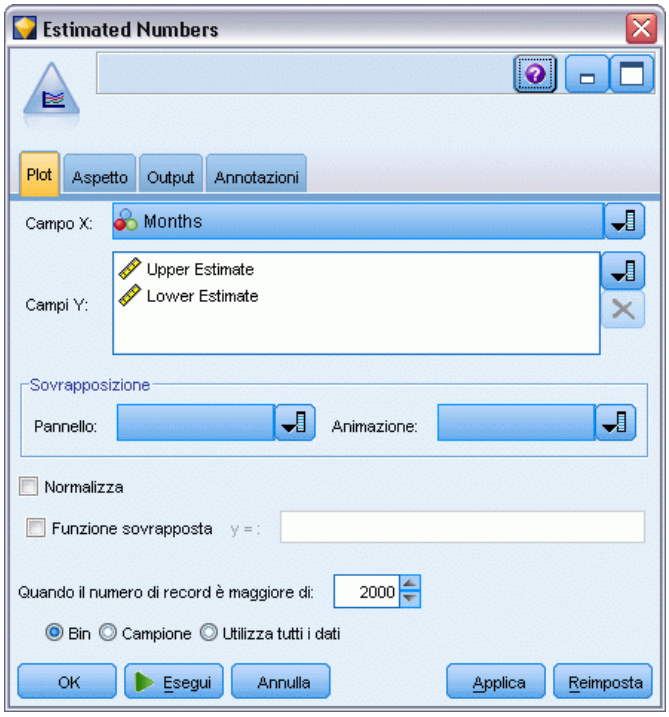
- Collegare un nodo Trasponi al nodo Filtro.
- Digitare 2 come numero di nuovi campi.

Figura 26-27
Nodo Filtro: scheda Filtro



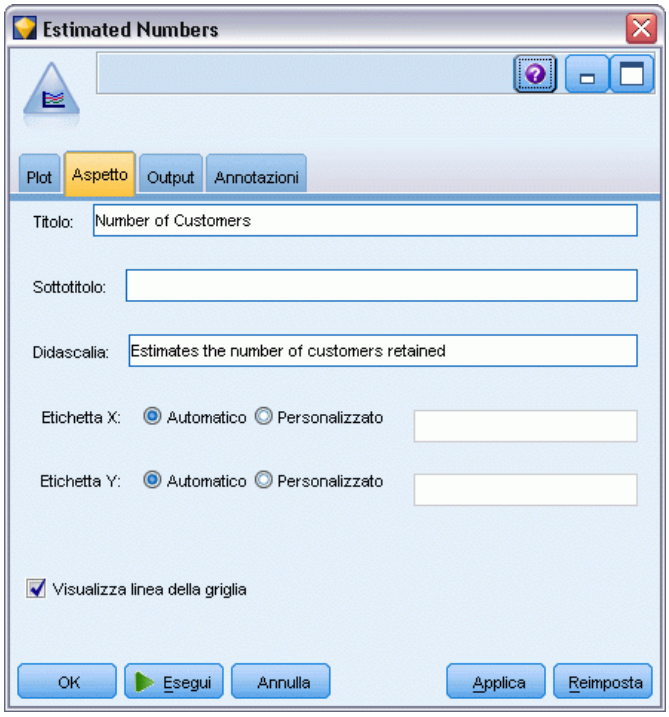
- Collegare un nodo Filtro al nodo Trasponi.
- Nella scheda Impostazioni del nodo Filtro, ridenominare *ID* in *Mesi*, *Field1* in *Stima inferiore* e *Field2* in *Stima superiore*.

Figura 26-28
Nodo Plot multiplo: scheda Plot



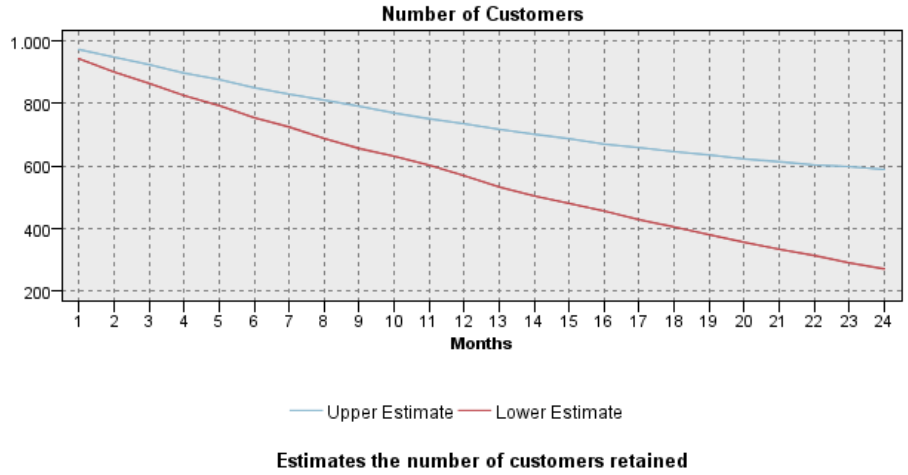
- Collegare un nodo Plot multiplo al nodo Filtro.
- Nella scheda Plot, specificare *Mesi* come campo X e *Stima inferiore* e *Stima superiore* come campi Y.

Figura 26-29
Nodo Plot multiplo: scheda Aspetto



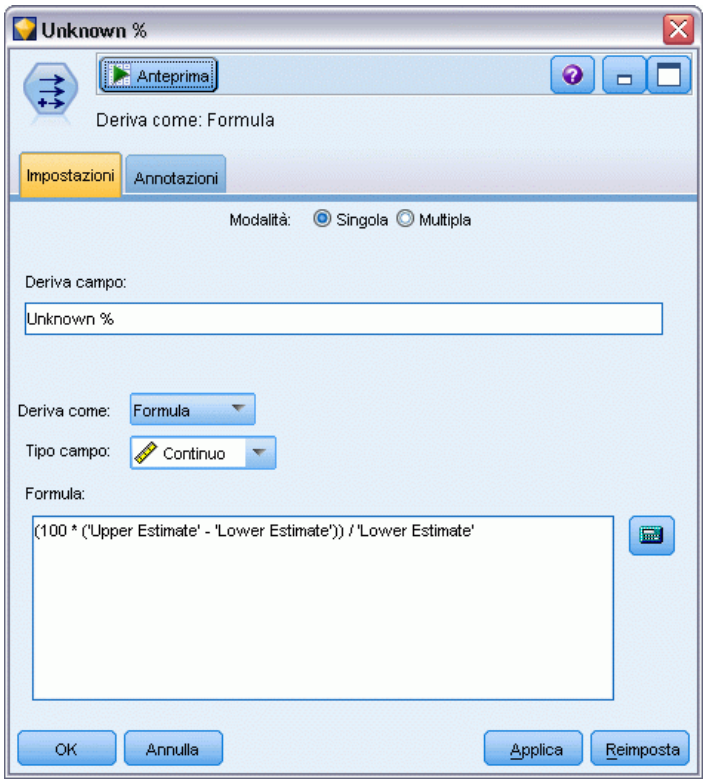
- Fare clic sulla scheda Aspetto.
- Digitare Numero di clienti come titolo.
- Digitare Stime del numero di clienti mantenuti come didascalia.
- Fare clic su Esegui.

Figura 26-30
Plot multiplo che stima il numero di clienti mantenuti



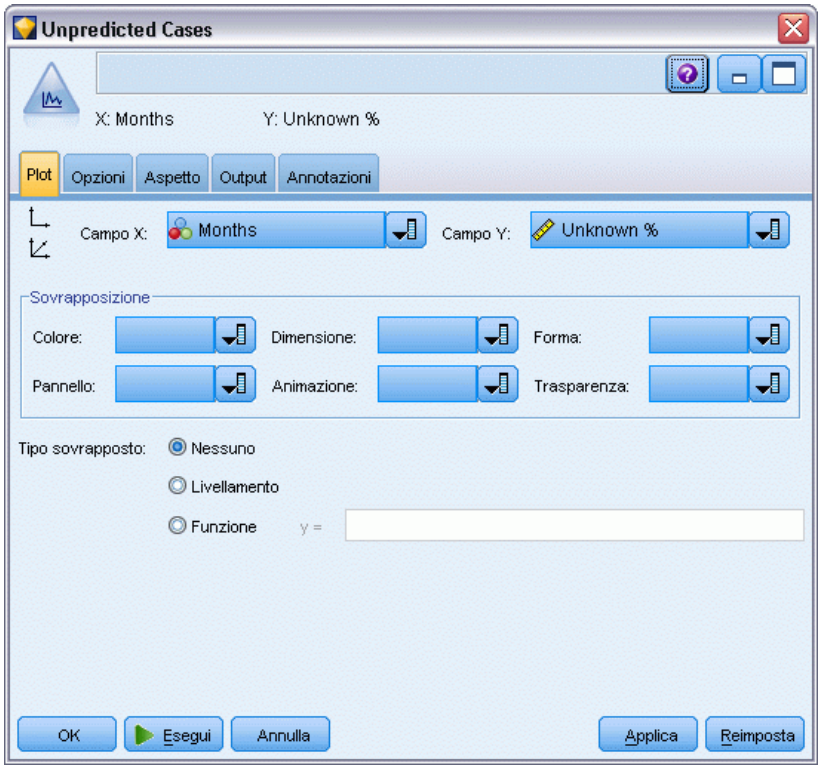
Viene rappresentato graficamente il limite inferiore e superiore del numero previsto di clienti mantenuti. La differenza tra le due linee rappresenta il numero di clienti il cui punteggio è stato calcolato come nullo e il cui stato è fortemente incerto. Nel tempo, il numero di questi clienti aumenta. Dopo 12 mesi, ci si può attendere di mantenere tra i 601 e i 735 dei clienti originali dell'insieme di dati, i quali dopo 24 mesi scendono tra i 288 e i 597.

Figura 26-31
Nodo Nuovo campo: scheda Impostazioni



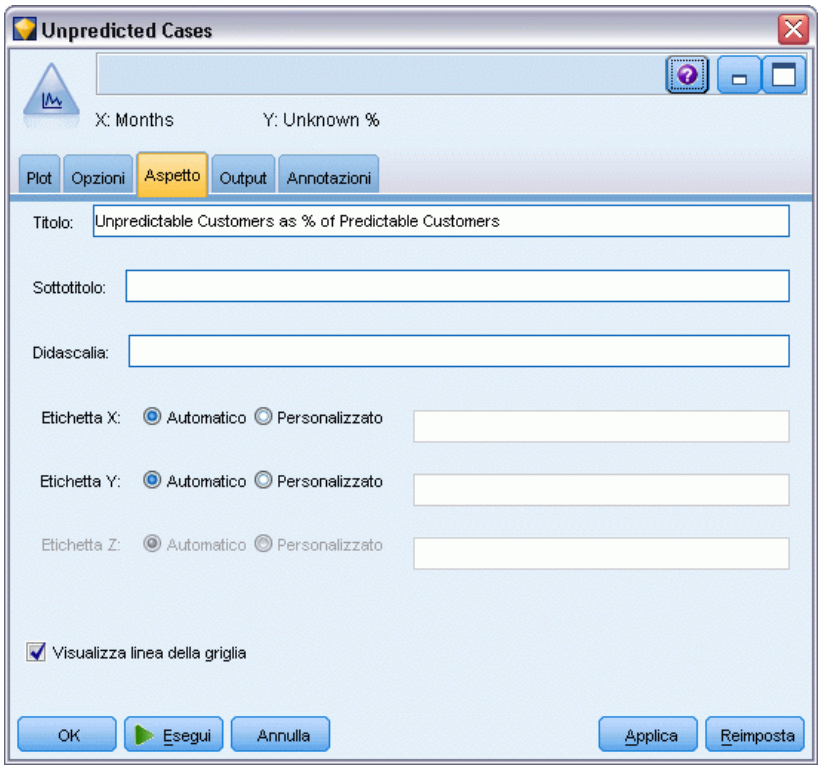
- Per analizzare ulteriormente il livello di incertezza delle stime del numero di clienti mantenuti, collegare un nodo Nuovo campo al nodo Filtro.
- Nella scheda Impostazioni del nodo Nuovo campo, digitare *% incerti* come nuovo campo.
- Selezionare Continuo come tipo di campo.
- Digitare (100 * ('Stima superiore' - 'Stima inferiore')) / 'Stima inferiore' come formula. *% incerti* è il numero di clienti "in dubbio", espresso come percentuale della stima inferiore.
- Fare clic su OK.

Figura 26-32
Nodo Plot: scheda Plot



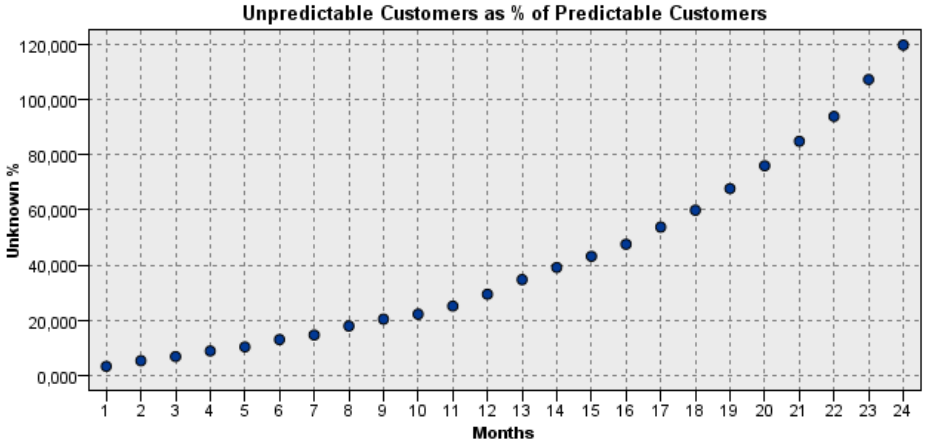
- Collegare un nodo Plot al nodo Nuovo campo.
- Nella scheda Plot del nodo Plot, selezionare *Mesi* come campo X e *% incerti* come campo Y.
- Fare clic sulla scheda Aspetto.

Figura 26-33
Nodo Plot: scheda Aspetto



- Digitare Clienti non prevedibili in % dei clienti prevedibili come titolo.
- Eseguire il nodo.

Figura 26-34
Rappresentazione grafica dei clienti non prevedibili

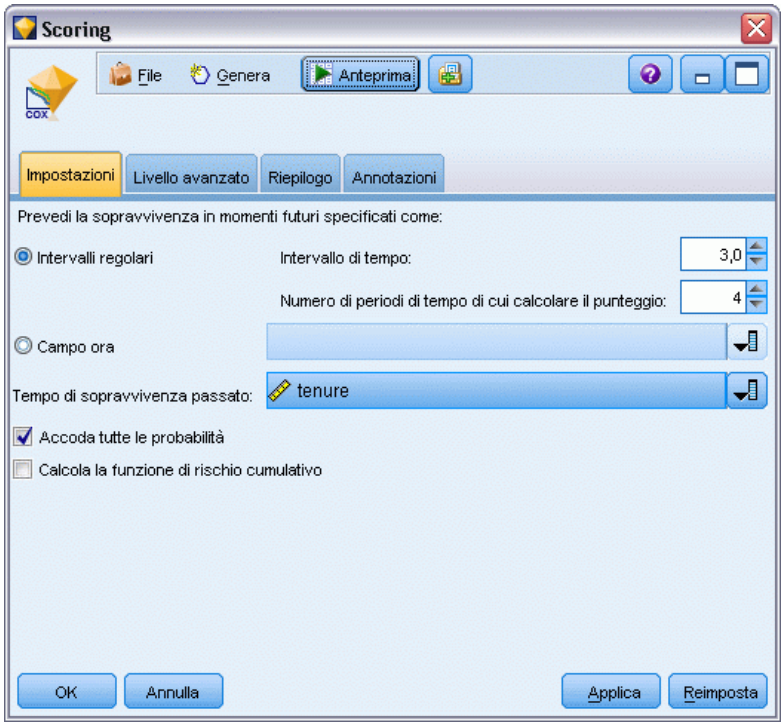


Nel corso del primo anno, la percentuale dei clienti non prevedibili aumenta pressoché linearmente, mentre il tasso di crescita esplode durante il secondo anno fino a che, in corrispondenza del mese 23, il numero di clienti con valori nulli supera il numero previsto di clienti mantenuti.

Calcolo del punteggio

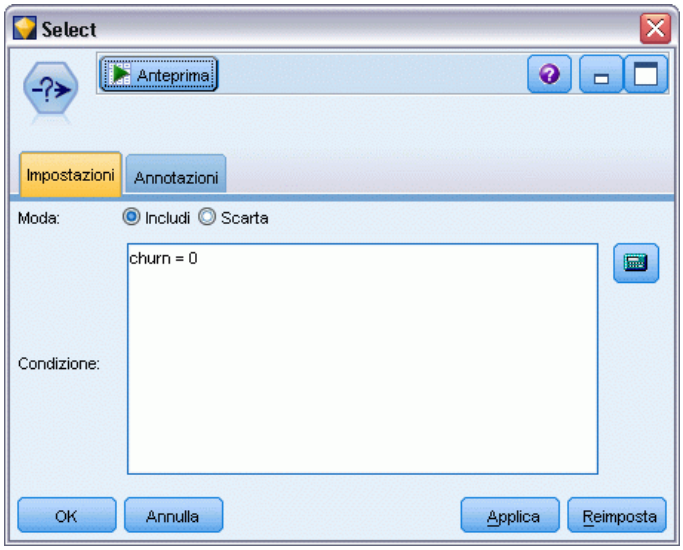
Una volta che si è soddisfatti del modello, si vorrà calcolare il punteggio dei clienti per identificare gli individui che hanno la maggior probabilità di passare a un altro operatore entro il prossimo anno, per trimestre.

Figura 26-35
Insieme di modelli Regressione di Cox: scheda Impostazioni



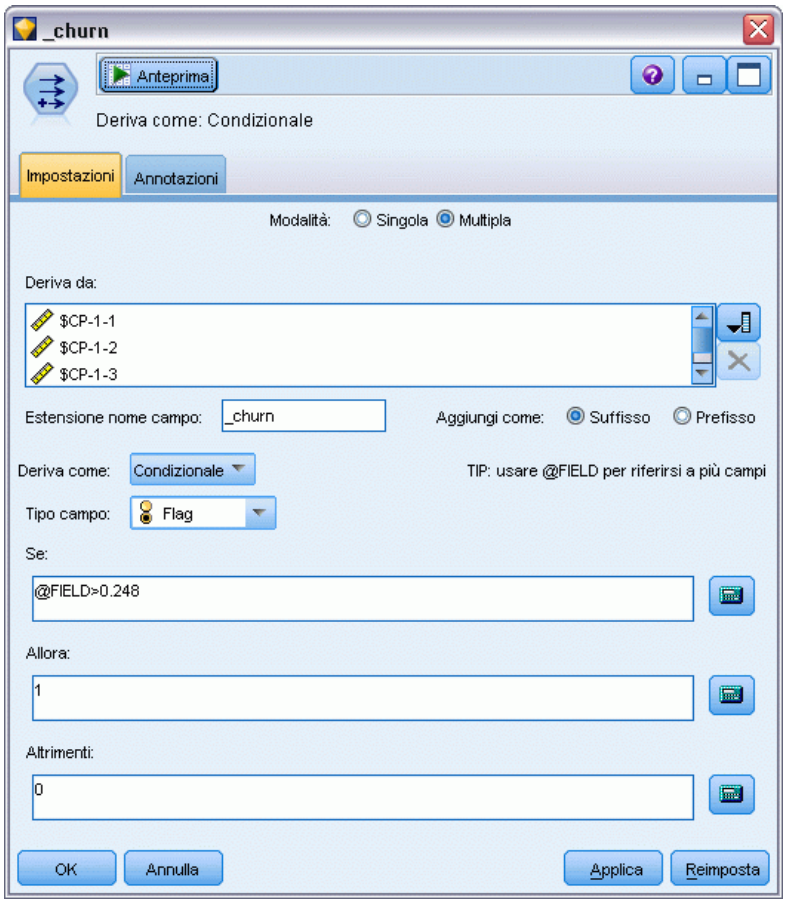
- Collegare un terzo insieme di modelli al nodo Input e aprire l'insieme di modelli.
- Assicurarsi che sia selezionata l'opzione Intervalli regolari e specificare 3.0 come intervallo temporale e 4 come numero di periodi di cui calcolare il punteggio. In questo modo, di ogni record verrà calcolato il punteggio per ognuno dei 4 trimestri successivi.
- Selezionare *tenure* come campo per specificare il tempo di sopravvivenza passato. L'algoritmo di calcolo del punteggio terrà conto della durata di ogni cliente come cliente della società.
- Selezionare Accoda tutte le probabilità. Questi campi aggiuntivi faciliteranno l'ordinamento dei record per la visualizzazione in una tabella.

Figura 26-36
Nodo Seleziona: scheda Impostazioni



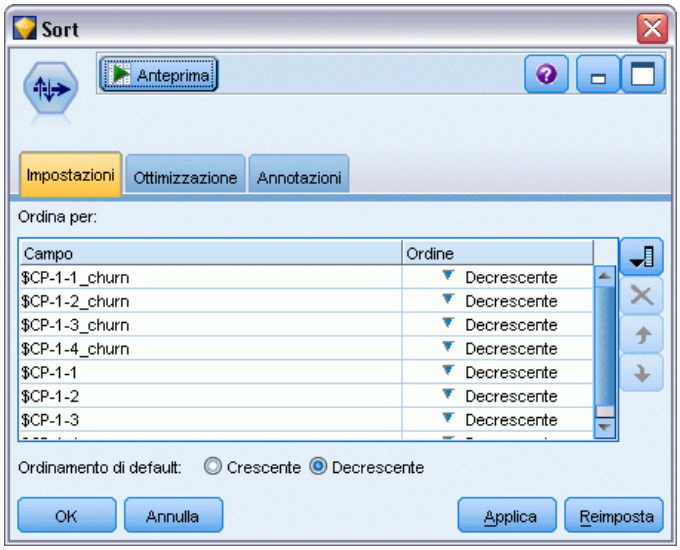
- Collegare un nodo Seleziona all'insieme di modelli. Nella scheda Impostazioni, digitare churn=0 come condizione. In questo modo dalla tabella dei risultati verranno rimossi i clienti che sono già passati a un altro operatore.

Figura 26-37
Nodo Nuovo campo: scheda Impostazioni



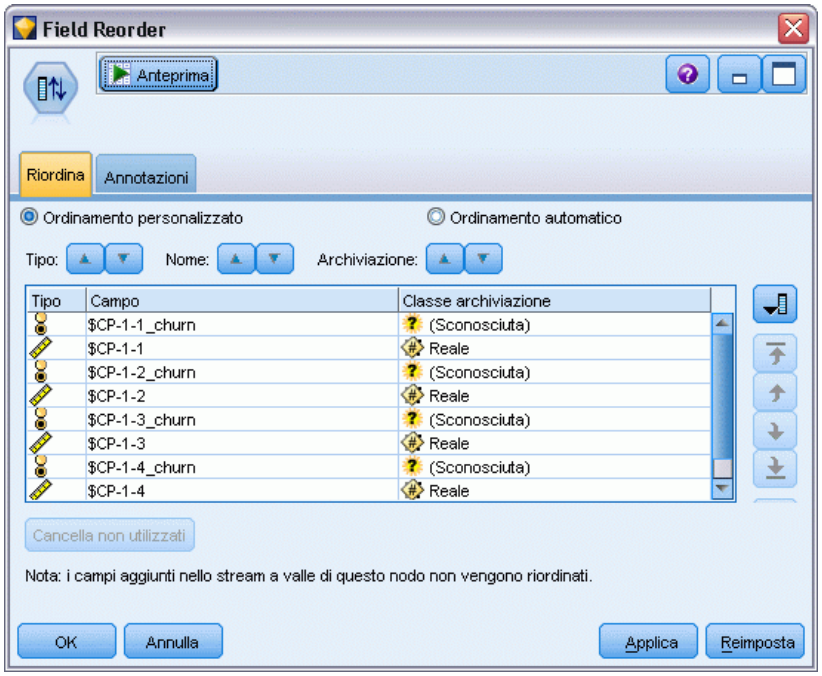
- Collegare un nodo Nuovo campo al campo Seleziona. Nella scheda Impostazioni, selezionare Multiplo come modalità.
- Scegliere di derivare dai campi da *$CP-1-1* a *$CP-1-4*, i campi con formato *$CP-1-n*, e digitare _abbandono come suffisso da aggiungere. Questa operazione risulta più semplice se nella finestra di dialogo Seleziona campi i campi vengono ordinati per Nome, ovvero in ordine alfabetico.
- Scegliere di derivare il campo come Condizionale.
- Selezionare Flag come livello di misurazione.
- Digitare @FIELD>0.248 come condizione Se. Si ricorda che questo valore era il valore di interruzione per la classificazione identificato mediante il nodo Valutazione.
- Digitare 1 come espressione Allora.
- Digitare 0 come espressione Altrimenti.
- Fare clic su OK.

Figura 26-38
Nodo Ordina: scheda Impostazioni



- collegare un nodo Ordina al nodo Nuovo campo. Nella scheda Impostazioni, scegliere di ordinare in base ai campi da *$CP-1-1_churn* e *$CP-1-4-churn* e quindi in base ai campi da *$CP-1-1* a *$CP-1-4*, tutto in ordine decrescente. I clienti che si prevede passeranno a un altro operatore appariranno in cima all'elenco.

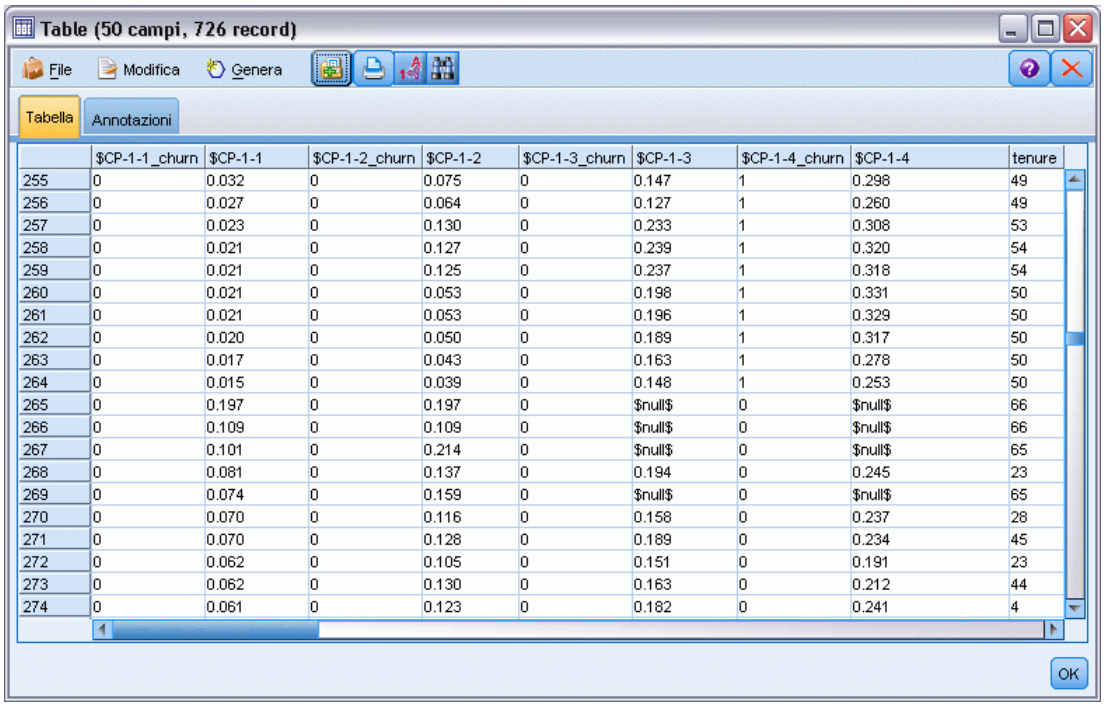
Figura 26-39
Nodo Riordina campi: scheda Riordina



- Collegare un nodo Riordina campi al nodo Ordina. Nella scheda Riordina, scegliere di posizionare i campi da *$CP-1-1_churn* a *$CP-1-4* prima degli altri. In questo modo la tabella dei risultati

risulterà più leggibile. Questa operazione è facoltativa. Per spostare i campi nella posizione indicata nella figura è necessario utilizzare i pulsanti.

Figura 26-40
Tabella che mostra i punteggi dei clienti



	$CP-1-1_churn	$CP-1-1	$CP-1-2_churn	$CP-1-2	$CP-1-3_churn	$CP-1-3	$CP-1-4_churn	$CP-1-4	tenure
255	0	0.032	0	0.075	0	0.147	1	0.298	49
256	0	0.027	0	0.064	0	0.127	1	0.260	49
257	0	0.023	0	0.130	0	0.233	1	0.308	53
258	0	0.021	0	0.127	0	0.239	1	0.320	54
259	0	0.021	0	0.125	0	0.237	1	0.318	54
260	0	0.021	0	0.053	0	0.198	1	0.331	50
261	0	0.021	0	0.053	0	0.196	1	0.329	50
262	0	0.020	0	0.050	0	0.189	1	0.317	50
263	0	0.017	0	0.043	0	0.163	1	0.278	50
264	0	0.015	0	0.039	0	0.148	1	0.253	50
265	0	0.197	0	0.197	0	$null$	0	$null$	66
266	0	0.109	0	0.109	0	$null$	0	$null$	66
267	0	0.101	0	0.214	0	$null$	0	$null$	65
268	0	0.081	0	0.137	0	0.194	0	0.245	23
269	0	0.074	0	0.159	0	$null$	0	$null$	65
270	0	0.070	0	0.116	0	0.158	0	0.237	28
271	0	0.070	0	0.128	0	0.189	0	0.234	45
272	0	0.062	0	0.105	0	0.151	0	0.191	23
273	0	0.062	0	0.130	0	0.163	0	0.212	44
274	0	0.061	0	0.123	0	0.182	0	0.241	4

- Collegare un nodo Tabella al nodo Riordina campi e avviarne l'esecuzione.

È previsto che 264 clienti passeranno a un altro operatore entro la fine dell'anno, 184 entro la fine del terzo trimestre, 103 entro il secondo e 31 nel primo. Si noti che, dati due clienti, quello con una propensione all'abbandono superiore nel primo trimestre non ha necessariamente una propensione all'abbandono superiore nei trimestri successivi. Per esempio, si considerino i record 256 e 260. Questo è probabilmente dovuto alla forma della funzione di rischio per i mesi successivi alla durata (tenure) corrente del cliente. Per esempio, potrebbe essere più probabile che i clienti che hanno scelto la società a seguito di una promozione passino a un altro operatore prima di quelli che l'hanno scelta su consiglio altrui, ma se non passano ad altro operatore in tale periodo potrebbero rimanere più fedeli di altri nella restante durata. Se necessario, è possibile riordinare i dati per ottenere visualizzazioni diverse dei clienti che hanno la maggiore probabilità di passare a un altro operatore.

Figura 26-41
Tabella che mostra i clienti con valori nulli

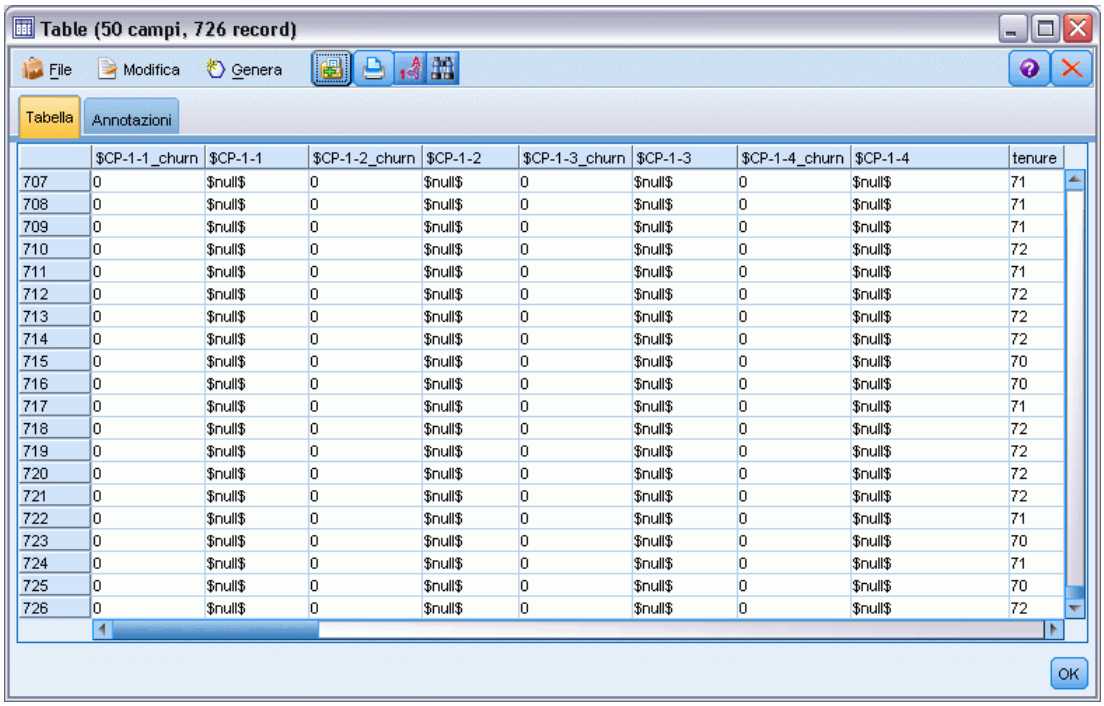


Table (50 campi, 726 record)

File Modifica Genera

Tabella Annotazioni

	$CP-1-1_churn	$CP-1-1	$CP-1-2_churn	$CP-1-2	$CP-1-3_churn	$CP-1-3	$CP-1-4_churn	$CP-1-4	tenure
707	0	$null$	0	$null$	0	$null$	0	$null$	71
708	0	$null$	0	$null$	0	$null$	0	$null$	71
709	0	$null$	0	$null$	0	$null$	0	$null$	71
710	0	$null$	0	$null$	0	$null$	0	$null$	72
711	0	$null$	0	$null$	0	$null$	0	$null$	71
712	0	$null$	0	$null$	0	$null$	0	$null$	72
713	0	$null$	0	$null$	0	$null$	0	$null$	72
714	0	$null$	0	$null$	0	$null$	0	$null$	72
715	0	$null$	0	$null$	0	$null$	0	$null$	70
716	0	$null$	0	$null$	0	$null$	0	$null$	70
717	0	$null$	0	$null$	0	$null$	0	$null$	71
718	0	$null$	0	$null$	0	$null$	0	$null$	72
719	0	$null$	0	$null$	0	$null$	0	$null$	72
720	0	$null$	0	$null$	0	$null$	0	$null$	72
721	0	$null$	0	$null$	0	$null$	0	$null$	72
722	0	$null$	0	$null$	0	$null$	0	$null$	71
723	0	$null$	0	$null$	0	$null$	0	$null$	70
724	0	$null$	0	$null$	0	$null$	0	$null$	71
725	0	$null$	0	$null$	0	$null$	0	$null$	70
726	0	$null$	0	$null$	0	$null$	0	$null$	72

OK

Nella parte inferiore della tabella sono elencati i clienti con valori nulli previsti. Si tratta dei clienti la cui durata totale (tempo futuro + *tenure*) non è compresa nell'intervallo dei tempi di sopravvivenza nei dati utilizzati per addestrare il modello.

Riepilogo

Utilizzando la regressione di Cox, è stato generato un modello accettabile del "tempo di abbandono", è stato rappresentato graficamente il numero previsto di clienti mantenuti nei successivi due anni e sono stati identificati i singoli clienti con la maggior probabilità di passare a un altro operatore nell'anno successivo. Si noti che, pur essendo questo un modello accettabile, potrebbe non essere quello migliore. Idealmente si dovrebbe quanto meno confrontare questo modello, ottenuto con il metodo stepwise in avanti, con uno creato con il metodo stepwise all'indietro.

Per una spiegazione dei fondamenti matematici dei metodi di modellazione utilizzati in IBM® SPSS® Modeler vedere il manuale *SPSS Modeler Algorithms Guide*.

Capitolo
27

Analisi market basket (Induzione di regole/C5.0)

In questo esempio vengono utilizzati dati fittizi che descrivono il contenuto dei basket (carrelli della spesa), ovvero insiemi di oggetti acquistati nella stessa circostanza, e i dati personali dell'acquirente, che possono essere acquisiti mediante uno schema del tipo "carta fedeltà". L'obiettivo è individuare gruppi di clienti che acquistano prodotti simili e che condividono le stesse caratteristiche demografiche, per esempio l'età, il reddito e così via.

Nell'esempio sono illustrate due fasi del data mining:

- Modelli di regole di associazione e una visualizzazione Web nella quale sono indicati i collegamenti tra gli oggetti acquistati
- Induzione della regola C5.0 per la creazione di profili degli acquirenti dei gruppi di prodotti specificati

Nota: In questa applicazione non viene utilizzata direttamente la creazione di modelli predittivi, pertanto il processo di data mining non prevede la misurazione della precisione dei modelli risultanti e nessuna distinzione delle fasi di addestramento e test associate.

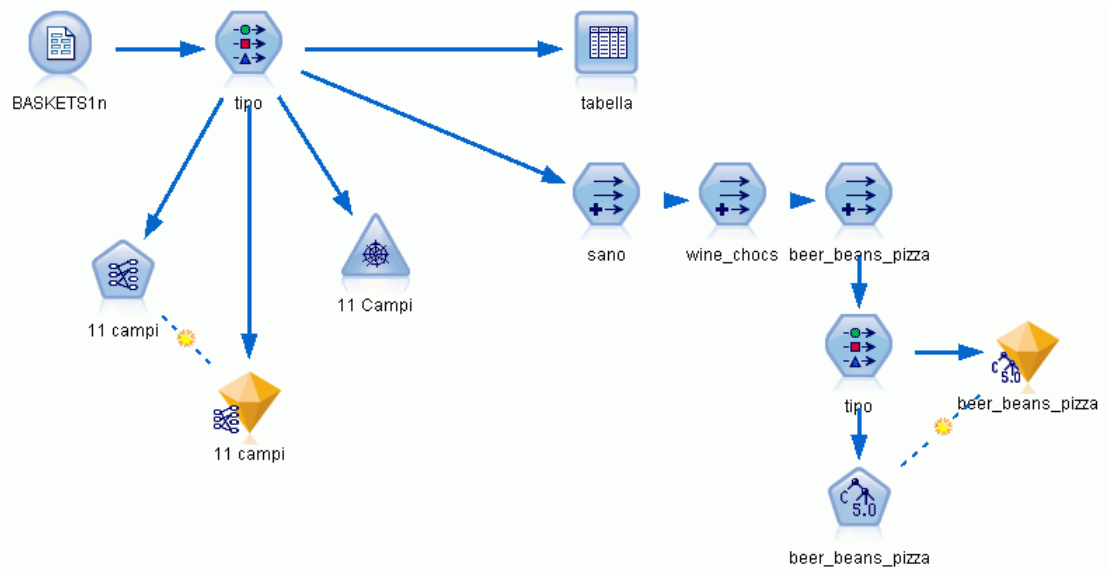
In questo esempio viene utilizzato lo stream denominato *baskrule* che fa riferimento al file di dati denominato *BASKETS1n*. Questi file sono disponibili nella directory *Demos* di tutte le installazioni di IBM® SPSS® Modeler. L'accesso si effettua dal gruppo di programmi IBM® SPSS® Modeler nel menu Start di Windows. Il file *baskrule* si trova nella directory *streams*.

Accesso ai dati

Utilizzare un nodo Testo variabile per connettersi all'insieme di dati *BASKETS1n*, specificando che i nomi dei campi verranno letti dal file. Connettere un nodo Tipo alla sorgente dati e quindi connettere il nodo a un nodo Tabella. Impostare il livello di misurazione del campo *id_carta* su *Senza tipo* (perché l'ID di ogni carta fedeltà è presente solo una volta nell'insieme di dati e pertanto non è utile per la creazione di modelli). Selezionare *Nominale* come livello di

misurazione del campo *sesso*, in modo da garantire che l'algoritmo per la creazione di modelli Apriori non consideri *sesso* come un flag.

Figura 27-1
Stream baskrule



Eseguire lo stream per istanziare il nodo Tipo e visualizzare la tabella. L'insieme di dati contiene 18 campi e ogni record rappresenta un basket.

I 18 campi vengono presentati con le seguenti intestazioni.

Riepilogo basket:

- *id_carta*. Identificatore della carta fedeltà per il cliente che acquista questo basket.
- *valore* Prezzo di acquisto totale del basket.
- *mod_pag*. Metodo di pagamento del basket.

Informazioni personali sul titolare della carta:

- *sesso*
- *casa_propria*. Specifica se il titolare della carta è o meno proprietario della propria abitazione.
- *reddito*
- *età*

Contenuto del basket—flag per la presenza di categorie di prodotti:

- *frutta_verdura*
- *carne*
- *latticini*
- *verdura_scatola*
- *carne_scatola*

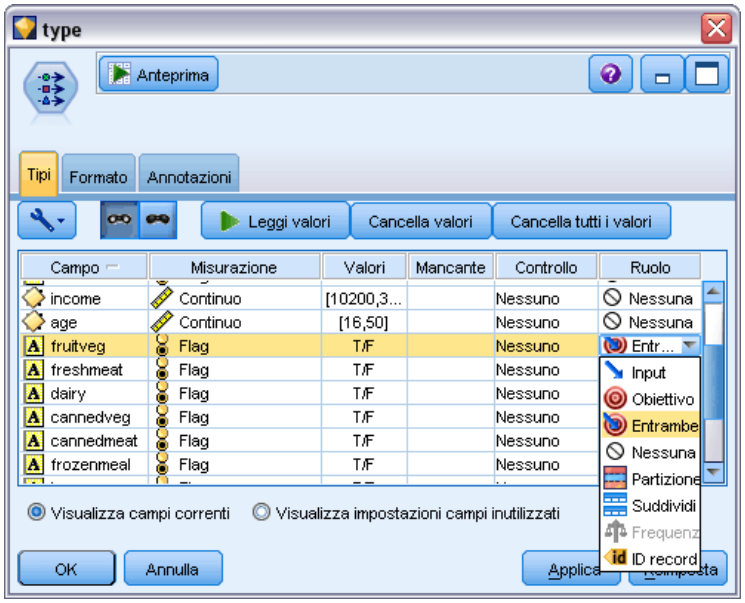
- *carne_surgelata*
- *birra*
- *vino*
- *softdrink*
- *pesce*
- *pasticceria*

Individuazione di affinità nel contenuto del basket

È necessario innanzitutto crearsi un'idea generale delle affinità (associazioni) nel contenuto del basket utilizzando Apriori per creare regole di associazione. Per selezionare i campi da utilizzare nel processo di creazione dei modelli, modificare il nodo Tipo e impostare il ruolo di tutte le categorie di prodotti su *Entrambi* e tutti gli altri ruoli su *Nessuno*. (Se si imposta *Entrambi*, è possibile che il campo sia un input o un output del modello risultante).

Nota: per impostare le opzioni relative a più campi, è possibile tenere premuto Maiusc e selezionare i campi prima di scegliere un'opzione dalle colonne.

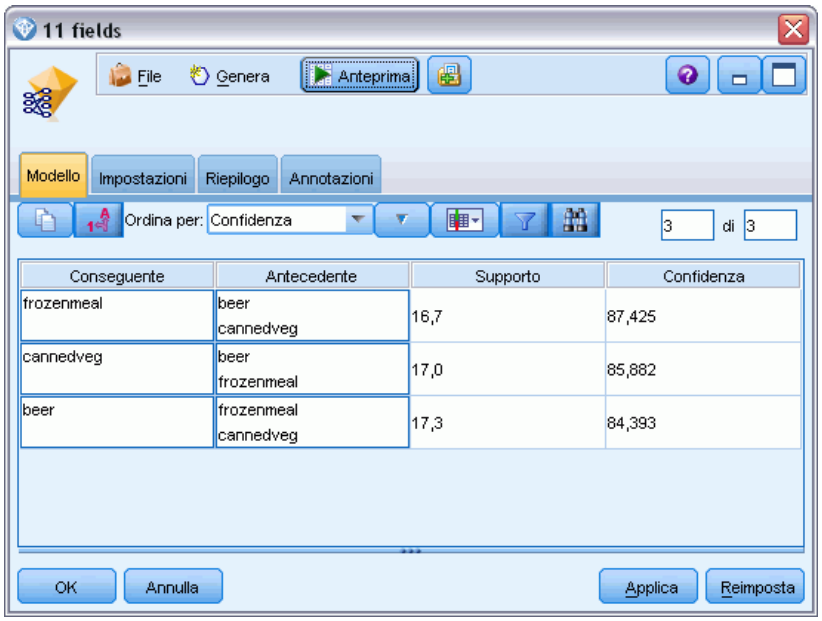
Figura 27-2
Selezione dei campi per la creazione di modelli



Dopo avere specificato i campi per la creazione di modelli, collegare un nodo Apriori al nodo Tipo, modificarlo, selezionare l'opzione Solo valori veri per i flag ed eseguire il nodo Apriori. Il risultato è un modello nella scheda Modelli disponibile in alto a destra nella finestra Manager.

Il modello contiene le regole di associazione che è possibile visualizzare mediante il menu di scelta rapida, scegliendo Visualizza.

Figura 27-3
Regole di associazione



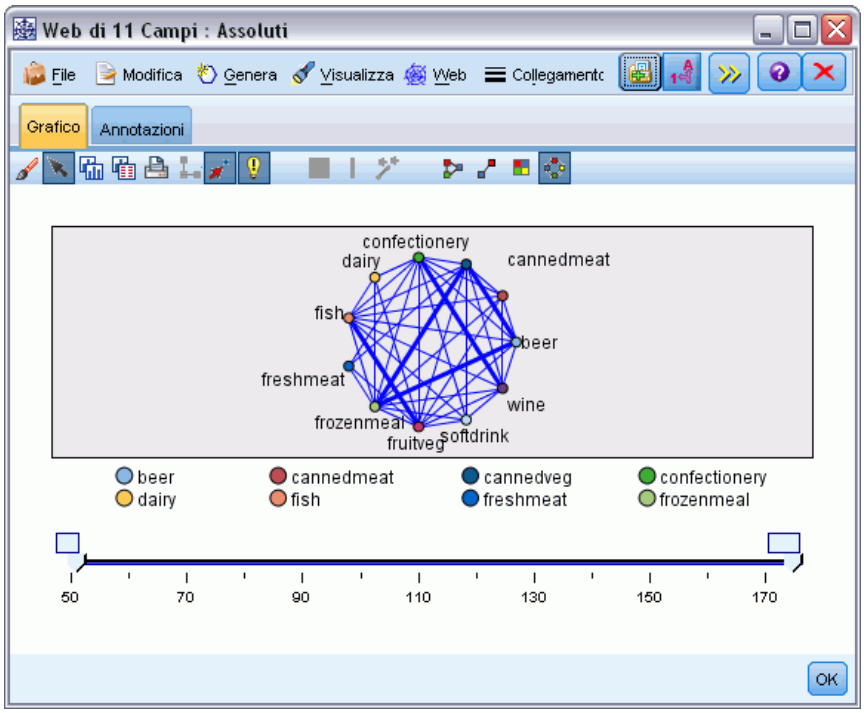
Le regole includono numerose associazioni tra carni surgelate, verdure in scatola e birra. La presenza di regole di associazioni a due vie quali:

```
frozenmeal -> beer
beer -> frozenmeal
```

suggerisce che una visualizzazione Web (nella quale vengono visualizzate solo associazioni a due vie) può evidenziare alcuni degli schemi dei dati.

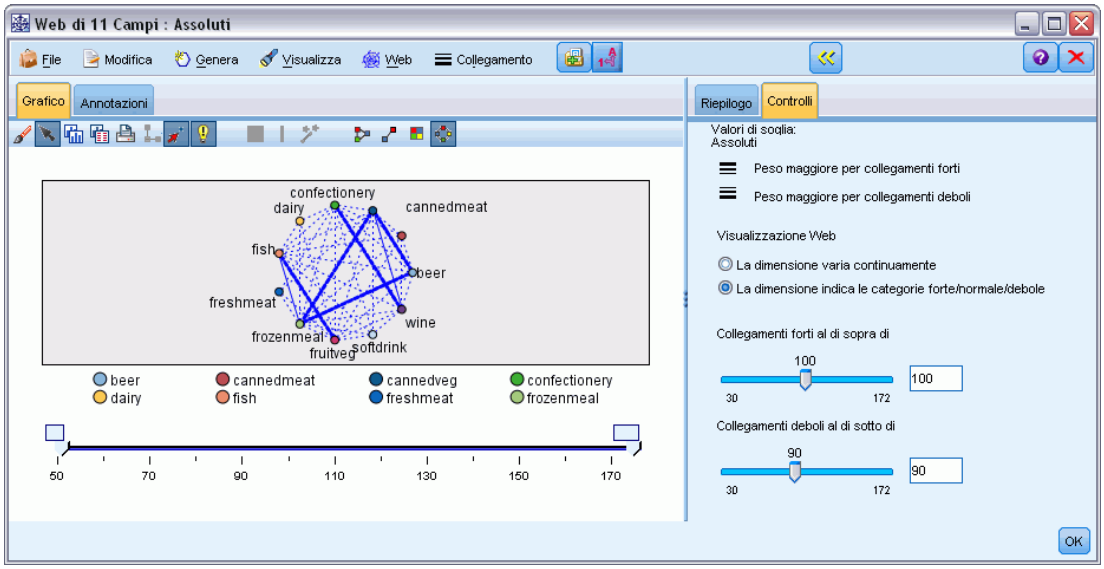
Collegare un nodo Web al nodo Tipo, modificare il nodo Web, selezionare tutti i campi del contenuto del basket, selezionare Mostra solo flag veri ed eseguire il nodo Web.

Figura 27-4
Visualizzazione Web delle associazioni di prodotti



Poiché la maggior parte delle combinazioni delle categorie di prodotti sono presenti in molti basket, i collegamenti forti in questo web sono troppi per visualizzare i gruppi di clienti suggeriti dal modello.

Figura 27-5
Visualizzazione Web ridotta



- Per specificare le connessioni deboli e forti, fare clic sulla doppia freccia gialla nella barra degli strumenti. Verrà espansa la finestra di dialogo in cui sono visualizzati il riepilogo e i controlli dell'output Web.
- Selezionare La dimensione indica le categorie forte/normale/debole.
- Impostare i collegamenti deboli al di sotto di 90.
- Impostare i collegamenti forti al di sopra di 100.

Nella visualizzazione risultante sono evidenziati tre gruppi di clienti:

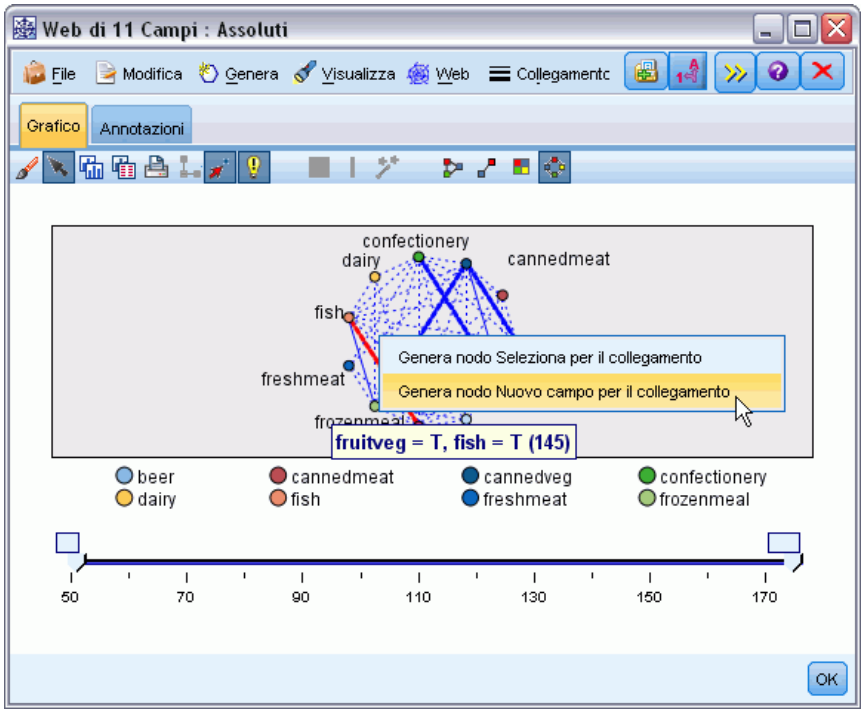
- I clienti che acquistano pesce e frutta e verdura
- I clienti che acquistano vino e pasticceria
- I clienti che acquistano birra, carni surgelate e verdure in scatola

Creazione di profili dei gruppi di clienti

A questo punto, sono stati identificati tre gruppi di clienti in base ai tipi di prodotti acquistati, ma è possibile anche avere maggiori informazioni su tali clienti, ovvero conoscerne il profilo demografico. A tale scopo, è possibile contrassegnare ogni cliente con un flag relativo a ognuno dei gruppi e utilizzare l'induzione di regola (C5.0) per creare profili basati sulla regola di tali flag.

È necessario innanzi tutto derivare un flag per ogni gruppo. Il flag può essere creato automaticamente mediante la visualizzazione Web appena creata. Con il pulsante destro del mouse, fare clic sul collegamento tra *frutta_verdura* e *pesce* per evidenziarlo, quindi selezionare con il pulsante destro del mouse Genera nodo Nuovo campo per il collegamento.

Figura 27-6
Derivazione di un flag per ogni gruppo di clienti



Modificare il nodo Nuovo campo risultante per modificare il nome del campo in *salute*. Ripetere l'operazione con il collegamento da *vino* a *pasticceria*, denominando il campo Nuovo campo risultante *vino_dolci*.

Per il terzo gruppo, che include tre collegamenti, assicurarsi innanzi tutto che non siano selezionati collegamenti. Fare clic tenendo premuto il tasto Maiusc per selezionare tutti e tre i collegamenti del triangolo *verdura_scatola*, *birra* e *carne_surgelata*, accertandosi di essere in modalità interattiva e non in modalità modifica. Quindi, dai menu della visualizzazione Web, scegliere:
Genera > Nuova variabile ("E")

Modificare il nome del campo Nuovo campo risultante in *birra_fagioli_pizza*.

Per creare il profilo di questi gruppi di clienti, connettere il nodo Tipo esistente ai tre nodi Nuovo campo in serie, quindi collegare un altro nodo Tipo. Nel nuovo nodo Tipo, impostare il ruolo *Nessuno* per tutti i campi a eccezione di *valore*, *mod_pag*, *sesso*, *casa_propria*, *reddito* ed *età*, per cui deve essere impostato *Input*, e per il gruppo di clienti rilevante (per esempio *birra_fagioli_pizza*), per cui deve essere impostato *Obiettivo*. Collegare un nodo C5.0, impostare il

tipo Output su Insieme di regole ed eseguire il nodo. Il modello risultante (per *birra_fagioli_pizza*) contiene un profilo demografico chiaro per il gruppo di clienti:

```
Rule 1 for T:
if sex = M
and income <= 16,900
then T
```

Lo stesso metodo può essere applicato ai flag dell'altro gruppo di clienti selezionandoli come output nel secondo nodo Tipo. Utilizzando Apriori anziché C5.0 in questo contesto è possibile creare una gamma più ampia di profili alternativi. Apriori consente inoltre di creare contemporaneamente profili di tutti i flag del gruppo di clienti, perché non è limitato a un singolo campo di output.

Riepilogo

In questo esempio è stato illustrato l'utilizzo di IBM® SPSS® Modeler per l'individuazione di affinità o collegamenti in un database, sia tramite la creazione di modelli (con Apriori) sia tramite la visualizzazione (con una visualizzazione Web). I collegamenti corrispondono a raggruppamenti di casi nei dati e per tali gruppi possono essere eseguite analisi dettagliate e creati profili mediante la creazione di modelli (con gli insiemi di regole C5.0).

Nell'ambito delle vendite al dettaglio, per esempio, i gruppi di clienti consentono di destinare le offerte speciali a target specifici, allo scopo di aumentare i tassi di risposta ai direct mailing o per personalizzare la gamma dei prodotti disponibili in una filiale in base alla domanda della specifica base demografica.

Capitolo
28

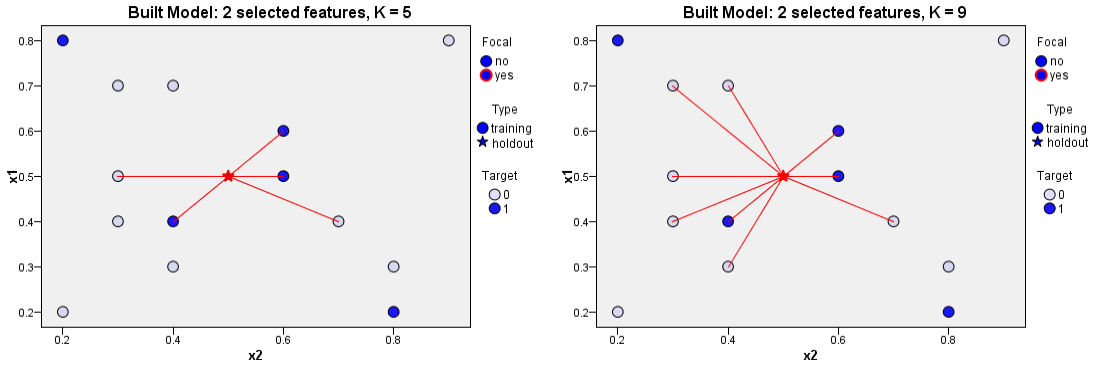
Valutazione di nuovi veicoli da commercializzare (KNN)

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 28-1
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.

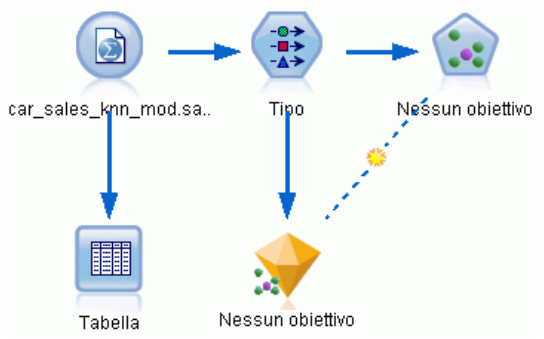
Una casa automobilistica ha sviluppato i prototipi di due nuovi veicoli, un'autovettura e un furgone. Prima di inserire i nuovi modelli nella sua linea di prodotti, l'azienda desidera determinare quali veicoli tra quelli già presenti sul mercato sono più simili ai prototipi, ovvero quali veicoli sono i loro "vicini più vicini", e quindi quali saranno i diretti concorrenti dei suoi due nuovi modelli.

Il produttore ha raccolto i dati relativi ai modelli esistenti suddividendoli in varie categorie e ha aggiunto i dettagli dei suoi prototipi. Le categorie secondo cui saranno confrontati i modelli comprendono prezzo in migliaia (*prezzo*), dimensioni del motore (*dim_motore*), potenza (*potenza*), interasse (*interasse*), larghezza (*larghezza*), lunghezza (*lunghezza*), peso a secco (*peso_asecco*), serbatoio carburante (*serb_carb*) e rendimento carburante (*mpg*).

In questo esempio viene utilizzato lo stream denominato *car_sales_knn.str*, disponibile nella sottocartella *streams* della cartella *Demos*. Il file di dati è *car_sales_knn_mod.sav*. Per ulteriori informazioni, vedere l'argomento Cartella Demos in il capitolo 1 in *Manuale dell'utente di IBM SPSS Modeler 15*.

Creazione dello stream

Figura 28-2
Stream di esempio per la modellazione KNN

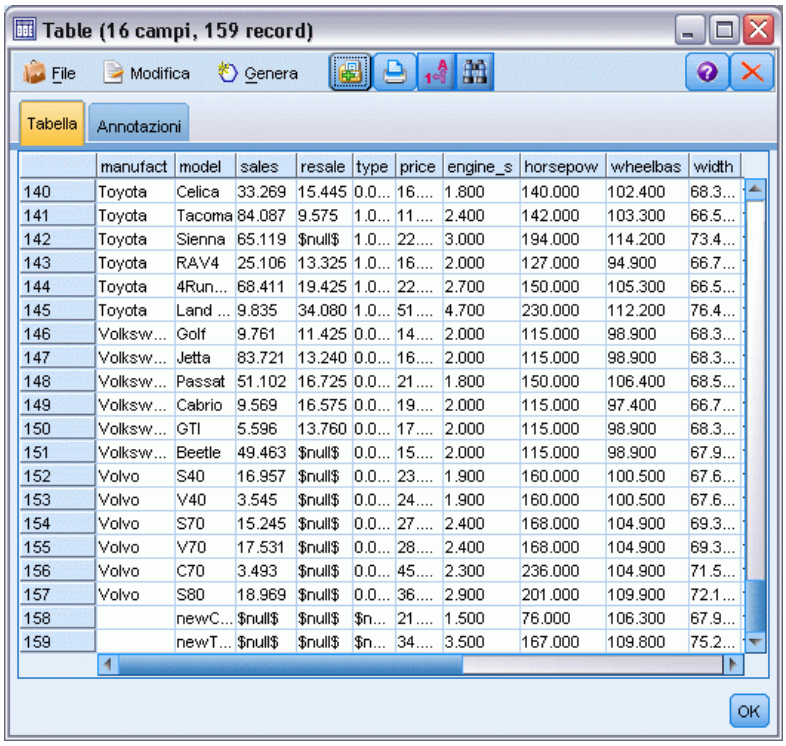


Creare un nuovo stream e aggiungere un nodo di input File Statistics che punti al file *car_sales_knn_mod.sav* situato nella directory *Demos* dell'installazione IBM® SPSS® Modeler.

Si osservino innanzitutto i dati raccolti dal produttore.

- Collegare un nodo Tabella al nodo di input File Statistics.
- Aprire il nodo Tabella e fare clic su Esegui.

Figura 28-3
Dati di origine per autovetture e furgoni



	manufact	model	sales	resale	type	price	engine_s	horsepow	wheelbas	width
140	Toyota	Celica	33.269	15.445	0.0...	16....	1.800	140.000	102.400	68.3...
141	Toyota	Tacoma	84.087	9.575	1.0...	11....	2.400	142.000	103.300	66.5...
142	Toyota	Sienna	65.119	$null$	1.0...	22....	3.000	194.000	114.200	73.4...
143	Toyota	RAV4	25.106	13.325	1.0...	16....	2.000	127.000	94.900	66.7...
144	Toyota	4Run...	68.411	19.425	1.0...	22....	2.700	150.000	105.300	66.5...
145	Toyota	Land ...	9.835	34.080	1.0...	51....	4.700	230.000	112.200	76.4...
146	Volksw...	Golf	9.761	11.425	0.0...	14....	2.000	115.000	98.900	68.3...
147	Volksw...	Jetta	83.721	13.240	0.0...	16....	2.000	115.000	98.900	68.3...
148	Volksw...	Passat	51.102	16.725	0.0...	21....	1.800	150.000	106.400	68.5...
149	Volksw...	Cabrio	9.569	16.575	0.0...	19....	2.000	115.000	97.400	66.7...
150	Volksw...	GTI	5.596	13.760	0.0...	17....	2.000	115.000	98.900	68.3...
151	Volksw...	Beetle	49.463	$null$	0.0...	15....	2.000	115.000	98.900	67.9...
152	Volvo	S40	16.957	$null$	0.0...	23....	1.900	160.000	100.500	67.6...
153	Volvo	V40	3.545	$null$	0.0...	24....	1.900	160.000	100.500	67.6...
154	Volvo	S70	15.245	$null$	0.0...	27....	2.400	168.000	104.900	69.3...
155	Volvo	V70	17.531	$null$	0.0...	28....	2.400	168.000	104.900	69.3...
156	Volvo	C70	3.493	$null$	0.0...	45....	2.300	236.000	104.900	71.5...
157	Volvo	S80	18.969	$null$	0.0...	36....	2.900	201.000	109.900	72.1...
158		newC...	$null$	$null$	$n...	21....	1.500	76.000	106.300	67.9...
159		newT...	$null$	$null$	$n...	34....	3.500	167.000	109.800	75.2...

In coda al file sono stati aggiunti i dettagli dei due prototipi, denominati *nuovaAuto* e *nuovoFurgone*.

Dai dati di origine si nota che il produttore utilizza la classificazione “furgone” (valore 1 nella colonna *tipo*) in modo piuttosto generico per indicare tutti i veicoli diversi dalle autovetture.

L’ultima colonna, *partizione*, è necessaria per far sì che i due prototipi possano essere definiti come record di controllo quando si tratterà di identificare i loro vicini più vicini. In questo modo i relativi dati non influiranno sui calcoli, poiché l’entità da considerare è il resto del mercato. Se si imposta il valore di *partizione* dei due record di controllo su 1 quando tutti gli altri record presentano un valore 0 in questo campo, esso può essere utilizzato in seguito al momento di impostare i record principali, cioè i record di cui si desidera calcolare i vicini più vicini.

Per il momento, lasciare aperta la finestra di output della tabella poiché si tornerà a esaminarla in seguito.

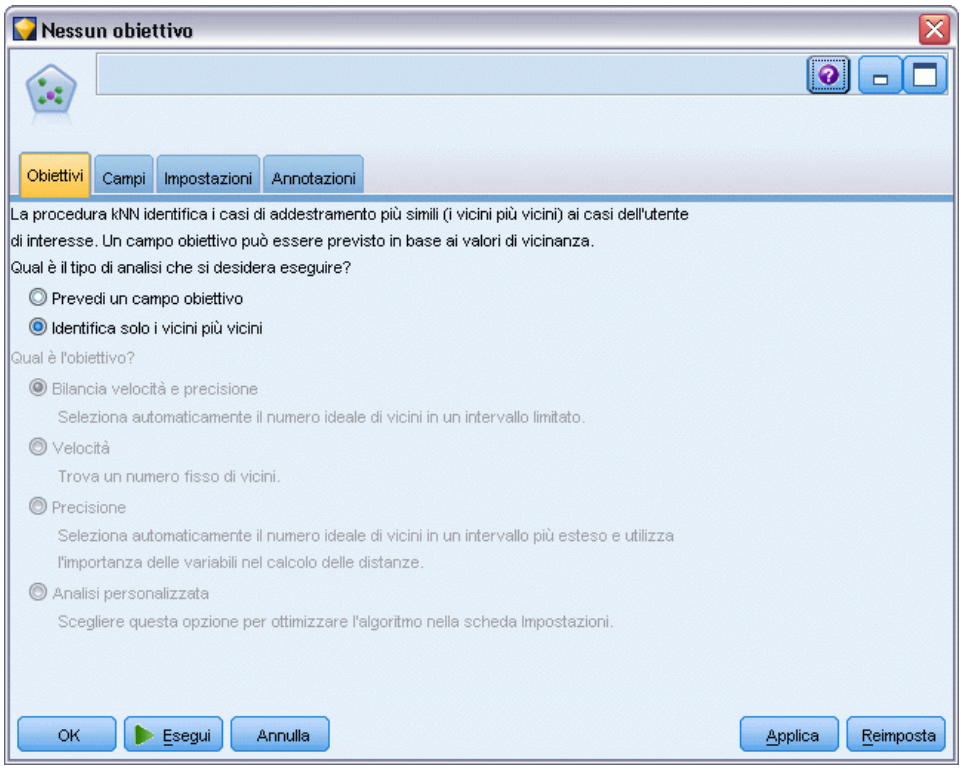
Figura 28-4
Impostazioni del nodo Tipo

- Aggiungere un nodo Tipo allo stream.
- Collegare il nodo Tipo al nodo di input File Statistics.
- Aprire il nodo Tipo.

  Dal momento che i campi da confrontare sono solo quelli compresi tra *prezzo* e *mpg*, lasciare il ruolo di tutti questi campi impostato su Input.

- Per tutti gli altri campi (da *produtt.* a *tipo*, più *Vendite*), impostare il ruolo su Nessuno.
- Impostare il livello di misurazione dell'ultimo campo, *partizione*, su Flag. Verificare che il relativo ruolo sia impostato su Input.
- Fare clic su Leggi valori per leggere i valori dei dati nello stream.
- Fare clic su OK.

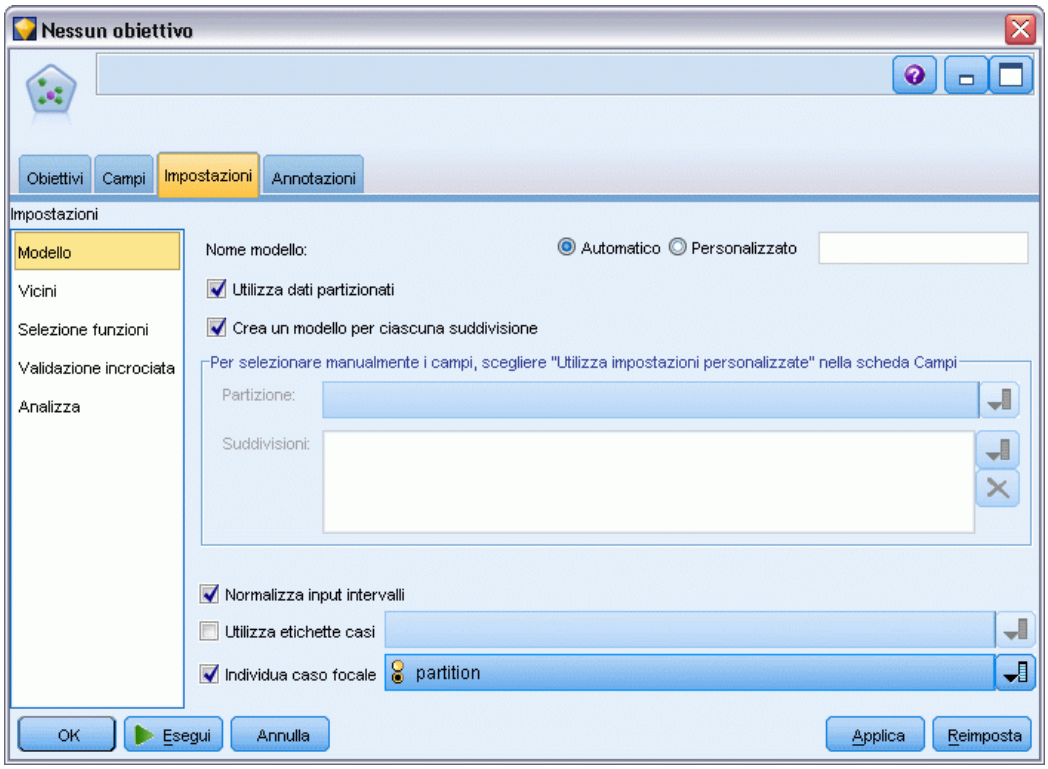
Figura 28-5
Scelta di identificare i vicini più vicini



- Collegare un nodo KNN al nodo Tipo.
- Aprire il nodo KNN.

  In questo caso non si effettuerà la previsione di un campo obiettivo, poiché lo scopo è semplicemente individuare i vicini più vicini dei prototipi in esame.
- Nella scheda Obiettivi, scegliere Identifica solo i vicini più vicini.
- Fare clic sulla scheda Impostazioni.

Figura 28-6
Utilizzo del campo partizione per individuare i record principali



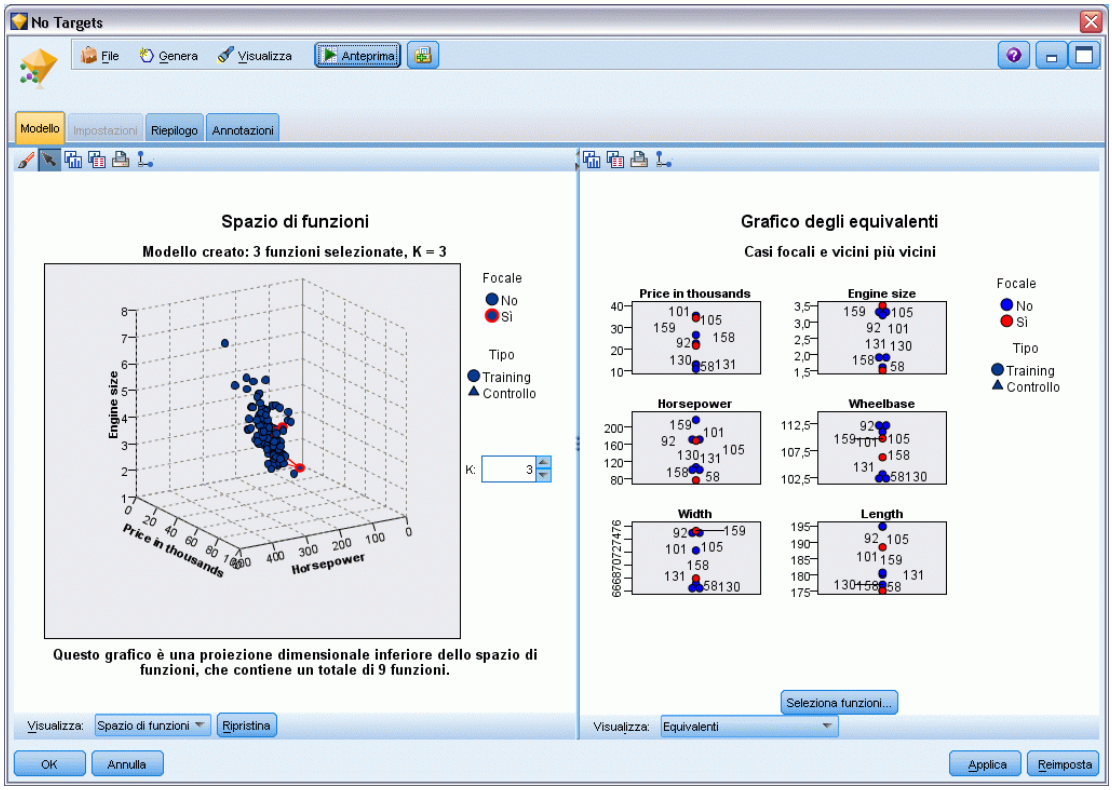
A questo punto è possibile utilizzare il campo *partizione* per individuare i record principali, cioè i record per cui si desidera individuare i vicini più vicini. L'impiego di un campo flag fa in modo che i record in cui il valore di questo campo è impostato su 1 diventino i record principali.

Come si è visto, gli unici record con valore 1 in questo campo sono *nuovaAuto* e *nuovoFurgone*: questi saranno dunque i record principali.

- Nel riquadro Modello della scheda Impostazioni, selezionare la casella di controllo Identifica record focale.
- Scegliere partizione dall'elenco a discesa di questo campo.
- Fare clic sul pulsante Esegui.

Esame del risultato

Figura 28-7
Finestra del Visualizzatore del modello



Nell'area di disegno dello stream e nella palette Modelli è stato creato un insieme di modelli. Aprire uno dei due insiemi di modelli per accedere al Visualizzatore del modello, la cui finestra è divisa in due riquadri:

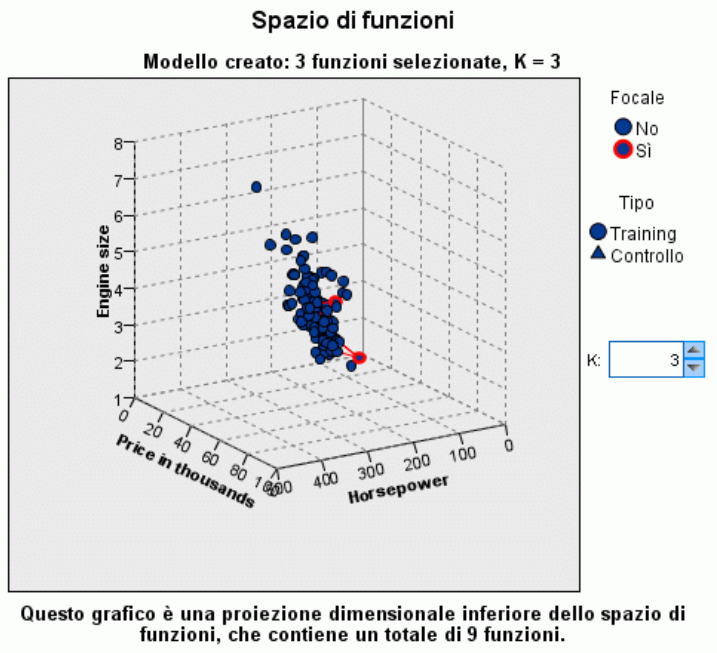
- Nel primo è presente una panoramica del modello denominata "vista principale". La vista principale del modello Vicino più vicino è detta **spazio di predittori**.
- Nel secondo, invece, possono essere visualizzate due tipologie di vista:

  La vista ausiliaria mostra ulteriori informazioni sul modello, pur non concentrandosi su quest'ultimo.

  La vista collegata mostra invece i dettagli relativi a una funzione del modello quando l'utente esegue il drill-down di parte della vista principale.

Spazio dei predittori

Figura 28-8
Grafico dello spazio di predittori



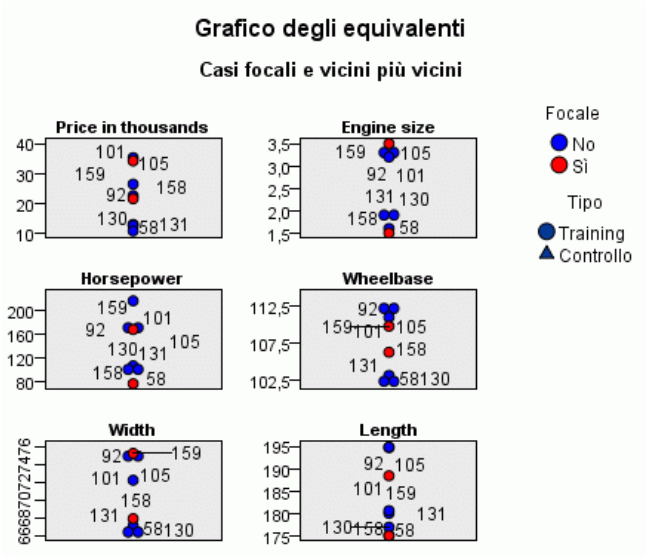
Il grafico dello spazio di predittori è un grafico tridimensionale interattivo che rappresenta i punti dei dati per tre funzioni (i primi tre campi di input dei dati di origine) che corrispondono a prezzo, dimensioni del motore e potenza.

I due record principali in oggetto sono evidenziati in rosso, con delle linee che li collegano ai rispettivi vicini più vicini *K*.

È possibile trascinare il grafico per ruotarlo e visualizzare più chiaramente la distribuzione dei punti nello spazio di predittori. Fare clic sul pulsante Ripristina per tornare alla visualizzazione di default.

Grafico degli equivalenti

Figura 28-9
Grafico degli equivalenti



La vista ausiliaria di default è il grafico degli equivalenti, che evidenzia i due record principali selezionati nello spazio di predittori e i relativi vicini più vicini *K* in base a sei singole funzioni (i primi sei campi di input dei dati di origine).

I veicoli sono rappresentati dai rispettivi numeri di record nei dati di origine. Per identificarli è necessario consultare l'output del nodo Tabella.

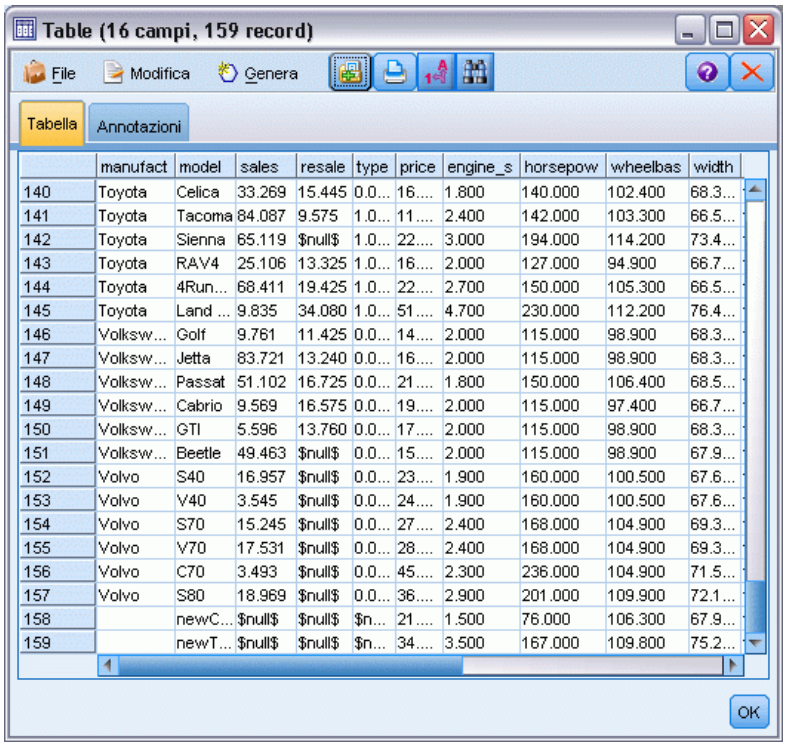
Se l'output del nodo Tabella è ancora disponibile:

- Fare clic sulla scheda Output nel riquadro dei manager nella parte superiore destra della finestra principale di IBM® SPSS® Modeler.
- Fare doppio clic sulla voce Tabella (16 campi, 159 record).

Se l'output del nodo Tabella non è più disponibile:

- Nella finestra principale di SPSS Modeler, aprire il nodo Tabella.
- Fare clic su Esegui.

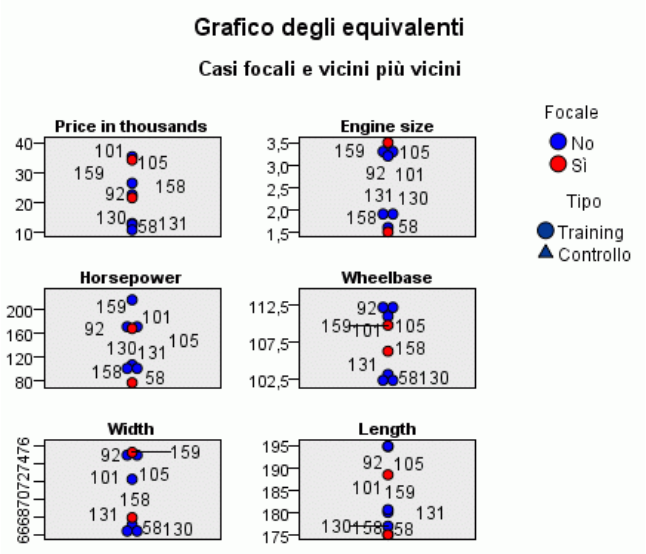
Figura 28-10
Individuazione dei record in base al numero



	manufact	model	sales	resale	type	price	engine_s	horsepow	wheelbas	width
140	Toyota	Celica	33.269	15.445	0.0...	16....	1.800	140.000	102.400	68.3...
141	Toyota	Tacoma	84.087	9.575	1.0...	11....	2.400	142.000	103.300	66.5...
142	Toyota	Sienna	65.119	$null$	1.0...	22....	3.000	194.000	114.200	73.4...
143	Toyota	RAV4	25.106	13.325	1.0...	16....	2.000	127.000	94.900	66.7...
144	Toyota	4Run...	68.411	19.425	1.0...	22....	2.700	150.000	105.300	66.5...
145	Toyota	Land ...	9.835	34.080	1.0...	51....	4.700	230.000	112.200	76.4...
146	Volksw...	Golf	9.761	11.425	0.0...	14....	2.000	115.000	98.900	68.3...
147	Volksw...	Jetta	83.721	13.240	0.0...	16....	2.000	115.000	98.900	68.3...
148	Volksw...	Passat	51.102	16.725	0.0...	21....	1.800	150.000	106.400	68.5...
149	Volksw...	Cabrio	9.569	16.575	0.0...	19....	2.000	115.000	97.400	66.7...
150	Volksw...	GTI	5.596	13.760	0.0...	17....	2.000	115.000	98.900	68.3...
151	Volksw...	Beetle	49.463	$null$	0.0...	15....	2.000	115.000	98.900	67.9...
152	Volvo	S40	16.957	$null$	0.0...	23....	1.900	160.000	100.500	67.6...
153	Volvo	V40	3.545	$null$	0.0...	24....	1.900	160.000	100.500	67.6...
154	Volvo	S70	15.245	$null$	0.0...	27....	2.400	168.000	104.900	69.3...
155	Volvo	V70	17.531	$null$	0.0...	28....	2.400	168.000	104.900	69.3...
156	Volvo	C70	3.493	$null$	0.0...	45....	2.300	236.000	104.900	71.5...
157	Volvo	S80	18.969	$null$	0.0...	36....	2.900	201.000	109.900	72.1...
158		newC...	$null$	$null$	$n...	21....	1.500	76.000	106.300	67.9...
159		newT...	$null$	$null$	$n...	34....	3.500	167.000	109.800	75.2...

Se si scorre la tabella fino in fondo si noterà che *nuovaAuto* e *nuovoFurgone* sono gli ultimi due record presenti nei dati, rispettivamente i numeri 158 e 159.

Figura 28-11
Confronto delle funzioni sul grafico degli equivalenti



Dal grafico degli equivalenti si nota, per esempio, che *nuovoFurgone* (159) ha una dimensione del motore maggiore rispetto ai suoi vicini più vicini, mentre *nuovaAuto* (158) ha un motore più piccolo rispetto a tutti i *suoi* vicini più vicini.

Per ciascuna delle sei funzioni è possibile spostare il mouse sopra i singoli punti per visualizzare il valore effettivo di ogni funzione in quel caso specifico.

Ma quali veicoli sono i vicini più vicini di *nuovaAuto* e *nuovoFurgone*?

Il grafico degli equivalenti contiene molti elementi ed è quindi opportuno passare a una visualizzazione più semplice.

- Fare clic sull'elenco a discesa Visualizza nella parte inferiore del grafico degli equivalenti (la voce visualizzata come Equivalenti).
- Selezionare Tabella dei vicini e delle distanze.

Tabella dei vicini e delle distanze

Figura 28-12
Tabella dei vicini e delle distanze

Vicini più vicini k e distanze

Visualizzata per casi focali iniziali

Caso focale	Vicini più vicini			Distanze più vicine		
	1	2	3	1	2	3
158	131	130	58	0,979	0,990	1,011
159	105	92	101	0,580	0,634	0,644

Questa visualizzazione è più chiara: si vedono infatti i tre modelli già presenti sul mercato a cui ciascuno dei due prototipi si avvicina di più.

Per *nuovaAuto* (record principale 158) i modelli sono Saturn SC (131), Saturn SL (130) e Honda Civic (58).

I risultati non sono particolarmente sorprendenti: tutte e tre sono berline di medie dimensioni, per cui *nuovaAuto* dovrebbe trovare una buona collocazione, soprattutto considerando il suo ottimo rendimento in termini di carburante.

Per *nuovoFurgone* (record principale 159), i vicini più vicini sono Nissan Quest (105), Mercury Villager (92) e Mercedes Classe M (101).

Come già visto, non si tratta necessariamente di furgoni in senso stretto, ma semplicemente veicoli classificati come diversi dalle autovetture. Se si osservano i vicini più vicini risultanti dall'elaborazione del nodo Tabella, si nota che *nuovoFurgone* ha un prezzo relativamente alto, oltre a essere uno dei veicoli più pesanti della sua categoria. Tuttavia, anche in questo caso il rendimento in termini di carburante è migliore rispetto ai più diretti concorrenti e questo dovrebbe essere un punto a suo favore.

Riepilogo

Nell'esempio precedente si è visto come utilizzare l'analisi del vicino più vicino per confrontare un insieme di funzioni molto diverse nei casi di un determinato insieme di dati. Sono stati inoltre calcolati, per due record di controllo molto diversi, i casi che più assomigliano ai record di controllo.

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.

Bibliografia

Asuncion, A., e D. Newman. 2007. "Repository per l'apprendimento automatico UCI." Available at http://mlearn.ics.uci.edu/MLRepository.html.

Indice