IBM SPSS Data Preparation 29

Nota

Prima di utilizzare queste informazioni e il prodotto che supportano, leggere le informazioni in "Informazioni particolari" a pagina 25.

Informazioni sul prodotto

Questa edizione si applica alla versione 29, release 0, modifica 1 di IBM® SPSS Statistics e a tutte le release e modifiche successive se non diversamente indicato nelle nuove edizioni.

Indice

Capitolo 1. Preparazione dei dati

Le seguenti funzioni di preparazione dei dati sono incluse nella Base Edition.

Introduzione alla preparazione dei dati

Poiché i sistemi di calcolo aumentano di potenza, gli appetiti per le informazioni crescono proporzionalmente, portando sempre più raccolta dati - più casi, più variabili e più errori di inserimento dati. Questi errori sono il bane delle previsioni del modello predittivo che sono l'obiettivo finale del data warehousing, quindi è necessario mantenere i dati "puliti". Tuttavia, la quantità di data warehoused è cresciuta finora oltre la possibilità di verificare i casi manualmente che è vitale implementare processi automatizzati per la validazione dei dati.

Data Preparazione consente di identificare casi insoliti e casi non validi, variabili e valori dati nel proprio dataset attivo e di preparare i dati per la modellazione.

Utilizzo delle procedure di preparazione dei dati

Il tuo utilizzo delle procedure di Data Preparazione dipende dalle tue particolari esigenze. Un itinerario tipico, dopo aver caricato i tuoi dati, è:

- **Preparazione dei metadati.** Esaminare le variabili nel proprio file dati e determinarne i valori validi, le etichette e i livelli di misurazione. Identificare combinazioni di valori variabili che sono impossibili ma comunemente sfumate. Definire le regole di convalida in base a queste informazioni. Questo può essere un compito dispendioso in termini di tempo, ma vale bene lo sforzo se è necessario convalidare periodicamente file di dati con attributi simili.
- **validazione dei dati.** Eseguire controlli di base e controlli contro le regole di convalida definite per identificare i casi non validi, le variabili e i valori dei dati. Quando vengono trovati dati non validi, indagare e correggere la causa. Questo può richiedere un altro passo attraverso la preparazione dei metadati.
- **Preparazione del modello.** Utilizzare la preparazione dei dati automatizzati per ottenere trasformazioni dei campi originali che miglioreranno l'edificio modello. Identificare potenziali outlier statistici che possono causare problemi per molti modelli predittivi. Alcuni outlier sono il risultato di valori variabili non validi che non sono stati identificati. Questo può richiedere un altro passo attraverso la preparazione dei metadati.

Una volta che il tuo file dati è "pulito", sei pronto a costruire modelli da altri moduli aggiuntivi.

Regole di convalida

Viene utilizzata una regola per determinare se un caso è valido. Esistono due tipi di regole di convalida:

- **Norme singole - variabili.** Le regole a singola variabile sono costituite da una serie fissa di controlli che si applicano ad una singola variabile, come ad esempio i controlli per i valori fuori gamma. Per le regole a singola variabile, i valori validi possono essere espressi come una serie di valori o un elenco di valori accettabili.
- **regole incrociate.** Le regole per più variabili sono regole definite dall'utente che possono essere applicate a singole variabili o a combinazioni di variabili. Le regole a variabile incrociata sono definite da un'espressione logica che indicava valori non validi.

Le regole di convalida vengono salvate nel dizionario dei dati del proprio file di dati. Questo consente di specificare una regola una volta e poi riutilizzarla.

Carica regole di convalida predefinite

È possibile ottenere rapidamente una serie di regole di convalida pronte all'uso caricando regole predefinite da un file di dati esterno incluso nell'installazione.

Per Caricare Regole Di Convalida Predefinite

1. Dai menu, scegliere:

   Dati > **Validazione** > **Carico Predefinito Regole ...**

In alternativa, è possibile utilizzare la procedura guidata di Copia delle proprietà dei dati per caricare le regole da qualsiasi file di dati.

Definisci regole di convalida

La Finestra di dialogo Define Validation Rules consente di creare e visualizzare regole di validazione a singola variabile e cross - variable.

Per creare e visualizzare le regole di convalida

1. Dai menu, scegliere:

   Dati > **Validazione** > **Definisci regole ...**

   La finestra di dialogo è popolata da regole di validazione a singola variabile e cross - variable lette dal dizionario dei dati. Quando non ci sono regole, viene creata automaticamente una nuova regola di segnaposto che puoi modificare per adattarti ai tuoi scopi.
2. Selezionare le singole regole sulle schede Regole singole e Cross - Variabili per visualizzare e modificare le relative proprietà.

Definire regole singole - variabili

La scheda Regole singole variabili consente di creare, visualizzare e modificare regole di convalida a singola variabile.

Regole. L'elenco riporta regole di validazione a singola variabile per nome e il tipo di variabile a cui la regola può essere applicata. Quando la finestra di dialogo viene aperta, mostra delle regole definite nel dizionario dei dati o, se non sono definite attualmente regole, una regola di segnaposto denominata "Single - Variable Rule 1". I seguenti pulsanti appaiono sotto l'elenco Regole:

- **Nuovo.** Aggiunge una nuova voce nella parte inferiore dell'elenco Regole. La regola è selezionata e viene assegnato il nome "SingleVarRule *n*," dove *n* è un intero in modo che il nome della nuova regola sia univoco tra le regole a variabile singolo e cross - variable.
- **Duplicato.** Aggiunge una copia della regola selezionata in fondo alla lista Regole. Il nome della regola viene regolato in modo che sia univoco tra le regole a variabile singolo e cross - variabile. Ad esempio, se si duplica "SingleVarregola 1", il nome della prima regola duplicata sarebbe "Copia di SingleVarregola 1", il secondo sarebbe "Copia (2) di SingleVarRule 1", e così via.
- **Elimina.** Elimina la regola selezionata.

Definizione di regola. Questi controlli consentono di visualizzare e impostare le proprietà per una regola selezionata.

- **Nome.** Il nome della regola deve essere unico tra le regole a singola variabile e cross - variable.
- **Tipo.** Questo è il tipo di variabile a cui la regola può essere applicata. Selezionare da **Numerico**, **String**e **Data**.
- **Formato.** Ciò consente di selezionare il formato data per le regole che possono essere applicate a variabili di data.
- **Valori validi.** È possibile specificare i valori validi sia come intervallo o un elenco di valori.

  Definizione di intervallo

I controlli di definizione di gamma consentono di specificare una gamma valida. I valori esterni all'intervallo sono contrassegnati come non validi.

Per specificare un intervallo, inserire i valori minimi o massimi, o entrambi. I controlli della casella di controllo consentono di contrassegnare valori non etichettati e non interi all'interno dell'intervallo.

Definizione di elenco

I controlli di definizione elenco consentono di definire un elenco di valori validi. I valori non inclusi nell'elenco sono contrassegnati come non validi.

Inserire i valori di elenco nella griglia. La casella di controllo determina se le questioni di caso quando i valori dei dati di stringa vengono controllati rispetto all'elenco dei valori accettabili.

- **Consenti valori mancanti utente.** Controlla se i valori mancanti dell'utente sono contrassegnati come non validi.
- **Consenti valori mancanti di sistema.** Controlla se i valori mancanti di sistema sono contrassegnati come non validi. Questo non si applica ai tipi di regola delle stringhe.
- **Consenti valori vuoti.** Controlla se i valori di stringa vuoti (cioè completamente vuoti) sono contrassegnati come non validi. Questo non si applica ai tipi di regola non stringa.

Definire regole cross - variabili

La scheda Regole Cross - Variabili consente di creare, visualizzare e modificare regole di convalida a variabili incrociate.

Regole. L'elenco mostra le regole di convalida incrociate per nome. Quando la finestra di dialogo viene aperta, mostra una regola di segnaposto chiamata "CrossVarRule 1". I seguenti pulsanti appaiono sotto l'elenco Regole:

- **Nuovo.** Aggiunge una nuova voce nella parte inferiore dell'elenco Regole. La regola è selezionata e viene assegnato il nome "CrossVarRule *n*," dove *n* è un numero intero in modo che il nome della nuova regola sia univoco tra le regole a variabile singolo e cross - variable.
- **Duplicato.** Aggiunge una copia della regola selezionata in fondo alla lista Regole. Il nome della regola viene regolato in modo che sia univoco tra le regole a variabile singolo e cross - variabile. Ad esempio, se si duplica "CrossVarRule 1", il nome della prima regola duplicata sarebbe "Copia di CrossVarRule 1", il secondo sarebbe "Copia (2) di CrossVarRule 1," e così via.
- **Elimina.** Elimina la regola selezionata.

Definizione di regola. Questi controlli consentono di visualizzare e impostare le proprietà per una regola selezionata.

- **Nome.** Il nome della regola deve essere unico tra le regole a singola variabile e cross - variable.
- **Espressione logica.** Si tratta, in sostanza, della definizione di regola. Si deve codificare l'espressione in modo che i casi non validi valutino a 1.

Creazione di espressioni

1. Per creare un'espressione, è possibile incollare o digitare direttamente i componenti nel campo Espressione.

- È possibile incollare funzioni o variabili di sistema comunemente utilizzate selezionando un gruppo dall'elenco del gruppo Funzione e facendo doppio clic sulla funzione o variabile nell'elenco Funzioni e Variabili speciali (oppure selezionare la funzione o la variabile e fare clic su **Inserisci**). Inserire valori per qualsiasi parametro indicato dai marchi di domanda (si applica solo alle funzioni). Il gruppo di funzioni etichettato **Tutto** fornisce un elenco di tutte le funzioni disponibili e le variabili di sistema. L'area dedicata alla finestra dialogo visualizza una breve descrizione della funzione o variabile correntemente selezionata.
- Le costanti stringa devono essere incluse tra virgolette o apostrofi.
- Se i valori contengono numeri decimali, è necessario utilizzare un punto (.) come indicatore decimale.

Convalida dati

La finestra di dialogo Dati di convalida consente di identificare casi sospetti e non validi, variabili e valori dati nel dataset attivo.

Esempio. Un analista di dati deve fornire una relazione mensile di soddisfazione del cliente al suo cliente. I dati che riceve ogni mese devono essere controllati di qualità per gli ID cliente incompleti, valori variabili fuori gamma e combinazioni di valori variabili comunemente inseriti in errore. La finestra di dialogo Dati Validate consente all'analista di specificare le variabili che identificano univocamente i clienti, definire regole a singola variabile per gli intervalli variabili validi e definire regole cross - variabili per catturare combinazioni impossibili. La procedura restituisce un report dei casi e delle variabili dei problemi. Inoltre, i dati hanno gli stessi elementi dati ogni mese, quindi l'analista è in grado di applicare le regole al nuovo file di dati il mese prossimo.

Statistiche. La procedura produce elenchi di variabili, casi e valori dati che fallisce vari controlli, conteggi di violazioni di regole a singola variabile e cross - variabili, e semplici riepiloghi descrittivi delle variabili di analisi.

Pesi. La procedura ignora la specifica della variabile di peso e la tratta invece come qualsiasi altra variabile di analisi.

Per convalidare i dati

1. Dai menu, scegliere:

   Dati > **Validazione** > **Validate dati ...**
2. Selezionare una o più variabili di analisi per la validazione mediante controlli variabili di base o con regole di convalida a singola variabile.

   In alternativa è possibile:
3. Fare clic sulla scheda **Regole cross - variabili** e applicare una o più regole cross - variabili.

Facoltativamente, è possibile:

- Selezionare una o più variabili di identificazione del caso per verificare gli ID duplicati o incompleti. Le variabili ID caso sono utilizzate anche per etichettare l'output casewise. Se vengono specificate due o più variabili ID caso, la combinazione dei loro valori viene trattata come identificativo del caso.

Campi con livello di misurazione sconosciuto

L'avviso Livello di misurazione viene visualizzato quando il livello di misurazione di una o più variabili (campi) del dataset è sconosciuto. Poiché influisce sul calcolo dei risultati di questa procedura, il livello di misurazione deve essere definito per tutte le variabili.

Esegui scansione dati. Legge i dati del dataset attivo e assegna un livello di misurazione predefinito a tutti i campi con livello di misurazione sconosciuto. Con dataset di grandi dimensioni, questa operazione può richiedere del tempo.

Assegna manualmente. Apre una finestra di dialogo che elenca tutti i campi con livello di misurazione sconosciuto, mediante la quale è possibile assegnare un livello di misurazione a questi campi. Il livello di misurazione si può assegnare anche nella Vista variabile dell'Editor dei dati.

Dal momento che il livello di misurazione è importante per questa procedura, è possibile accedere alla finestra di dialogo per la sua esecuzione solo quando per tutti i campi è stato definito un livello di misurazione.

Convalida Controlli di base dati

La Scheda Controlli di base consente di selezionare i controlli di base per le variabili di analisi, gli identificativi dei casi e i casi interi.

Variabili di analisi. Se sono state selezionate delle variabili di analisi sulla scheda Variabili, è possibile selezionare uno qualsiasi dei seguenti controlli di validità. La casella di controllo consente di attivare o disattivare i controlli.

- **Percentuale massima dei valori mancanti.** Riporta le variabili di analisi con una percentuale di valori mancanti superiore al valore specificato. Il valore specificato deve essere un numero positivo inferiore o uguale a 100.
- **La percentuale massima di casi in una singola categoria.** Se le eventuali variabili di analisi sono categoriali, questa opzione riporta variabili di analisi categoriali con una percentuale di casi che rappresenta una singola categoria non mancante superiore al valore specificato. Il valore specificato deve essere un numero positivo inferiore o uguale a 100. La percentuale si basa su casi con valori non mancanti della variabile.
- **Percentuale massima di categorie con conteggio di 1.** Se le eventuali variabili di analisi sono categoriali, questa opzione riporta variabili di analisi categoriali in cui la percentuale delle categorie della variabile contenente un solo caso è maggiore del valore specificato. Il valore specificato deve essere un numero positivo inferiore o uguale a 100.
- **Coefficiente di variazione minimo.** Se le eventuali variabili di analisi sono in scala, questa opzione riporta variabili di analisi di scala in cui il valore assoluto del coefficiente di variazione è inferiore al valore specificato. Questa opzione si applica solo alle variabili in cui la media è nonzero. Il valore specificato deve essere un numero non negativo. Specificando 0 disattiva il controllo del coefficiente di variazione.
- **Deviazione standard minima.** Se le eventuali variabili di analisi sono in scala, questa opzione riporta le variabili di analisi di scala la cui deviazione standard è inferiore al valore specificato. Il valore specificato deve essere un numero non negativo. Specificando 0 disattiva il controllo di deviazione standard.

Identificativi di causa. Se sono state selezionate delle variabili identificativo del caso sulla scheda Variabili, è possibile selezionare uno qualsiasi dei seguenti controlli di validità.

- **Flag ID incompleti.** Questa opzione riporta i casi con identificativi di casi incompleti. Per un caso particolare, un identificativo è considerato incompleto se il valore di qualsiasi variabile ID è vuoto o mancante.
- **Flag duplicati ID.** Questa opzione riporta i casi con identificativi di case duplicati. Gli identificativi incompleti sono esclusi dal set di possibili duplicati.

Flag vuoto. Questa opzione riporta i casi in cui tutte le variabili sono vuote o vuote. Allo scopo di identificare i casi vuoti, è possibile scegliere di utilizzare tutte le variabili presenti nel file (ad eccezione di eventuali variabili ID) o solo variabili di analisi definite sulla scheda Variabili.

Convalida dati Single - Regole variabili

La scheda Regole singole variabili visualizza le regole di validazione a singola variabile disponibili e consente di applicarle alle variabili di analisi. Per definire ulteriori regole a singola variabile, fare clic su **Definisci regole**. Per ulteriori informazioni, consultare la sezione “Definire regole singole - variabili” a pagina 2.

Variabili di analisi. L'elenco mostra le variabili di analisi, riepiloga le loro distribuzioni e mostra il numero di regole applicate ad ogni variabile. Si noti che i valori mancanti dell'utente e del sistema non sono inclusi nei riepiloghi. Il Display drop-down elenco controlla quali variabili vengono mostrate; è possibile scegliere tra **Tutte le variabili**, **Variazioni numeriche**, **Variabili di stringa**e **Variabili di data**.

Regole. Per applicare le regole alle variabili di analisi, selezionare una o più variabili e controllare tutte le regole che si desidera applicare nell'elenco Regole. L'elenco Regole mostra solo regole appropriate per le variabili di analisi selezionate. Ad esempio, se vengono selezionate variabili di analisi numerica, vengono mostrate solo regole numeriche; se viene selezionata una variabile stringa, vengono mostrate solo regole di stringa. Se non vengono selezionate variabili di analisi o hanno tipi di dati misti, non vengono mostrate regole.

Distribuzioni variabili. I riepiloghi di distribuzione mostrati nell'elenco Variabili di analisi possono essere basati su tutti i casi o su una scansione dei primi casi *n* , come specificato nella casella di testo Casi. Cliccando su **Rescan** si aggiorna i riepiloghi di distribuzione.

Validata Data Cross - Regole variabili

La scheda Regole cross - variabili visualizza le regole cross - variabili disponibili e consente di applicarle ai tuoi dati. Per definire ulteriori regole cross - variabili, fare clic su **Definisci regole**. Per ulteriori informazioni, consultare la sezione “Definire regole cross - variabili” a pagina 3.

Convalida output dati

Report Casewise. Se hai applicato regole di convalida a singola variabile o cross - variable, è possibile richiedere un report che elenchi le violazioni delle regole di convalida per singoli casi.

- **Numero minimo di violazioni.** Questa opzione specifica il numero minimo di violazioni delle regole necessarie per un caso da inserire nel report. Specificare un intero positivo.
- **Numero massimo di casi.** Questa opzione specifica il numero massimo di casi inclusi nel report del caso. Specificare un intero positivo inferiore o uguale a 1000.

Regole di convalida a singola variabile. Se avete applicato delle regole di convalida a singola variabile, è possibile scegliere come visualizzare i risultati o se visualizzarli a tutti.

- **Sintetizza le violazioni per variabile di analisi.** Per ogni variabile di analisi, questa opzione mostra tutte le regole di validazione a singola variabile violate e il numero di valori che violavano ogni regola. Segnala inoltre il numero totale di violazioni delle regole a singola variabile per ogni variabile.
- **Sintetizza le violazioni per regola.** Per ogni regola di convalida a singola variabile, questa opzione riporta le variabili che violano la regola e il numero di valori non validi per variabile. Riporta anche il numero totale di valori che violavano ogni regola attraverso le variabili.

Visualizza le statistiche descrittive per le variabili di analisi. Questa opzione consente di richiedere statistiche descrittive per le variabili di analisi. Viene generata una tabella di frequenza per ogni variabile categoriale. Per le variabili di scala è generata una tabella di statistiche di riepilogo tra cui la media, la deviazione standard, minima e massima.

Sposta i casi con violazioni delle regole di convalida ai vertici del dataset attivo. Questa opzione sposta i casi con violazioni di regola a singola variabile o cross - variabile alla parte superiore del dataset attivo per un facile perlettura.

Convalida dati Data salvataggio

La scheda Salva consente di salvare le variabili che registrano le violazioni delle regole al dataset attivo.

Variabili di riepilogo. Si tratta di variabili individuali che possono essere salvate. Controllare una casella per salvare la variabile. Vengono forniti i nomi predefiniti per le variabili; è possibile modificarli.

- **Indicatore di caso vuoto.** Ai casi vuoti viene assegnato il valore 1. Tutti gli altri casi sono codificati 0. I valori della variabile riflettono l'ambito specificato sulla scheda Controlli di Base.
- **Gruppo ID duplicato.** I casi che hanno lo stesso identificativo del caso (diverso dai casi con identificativi incompleti) vengono assegnati lo stesso numero di gruppo. I casi con identificativi unici o incompleti sono codificati 0.
- **Indicatore ID incompleto.** Ai casi con identificativi di caso vuoti o incompleti viene assegnato il valore 1. Tutti gli altri casi sono codificati 0.
- **Violazioni delle regole di convalida.** Questo è il conteggio totale casewise delle violazioni delle regole di convalida a singola variabile e cross - variable.

Sostituire le variabili di riepilogo esistenti. Le variabili salvate nel file dati devono avere nomi univoli o sostituire le variabili con lo stesso nome.

Variabili indicatori di salvataggio. Questa opzione consente di salvare un record completo di violazioni delle regole di convalida. Ogni variabile corrisponde a un'applicazione di una regola di convalida e ha un valore di 1 se il caso viola la regola e un valore di 0 se non lo fa.

Preparazione automatica dati

La preparazione dei dati per l'analisi rappresenta una delle fasi più importanti in qualsiasi progetto — e, tradizionalmente, una delle attività che richiedono più tempo. La funzione Preparazione automatica dati (ADP) svolge questo compito al posto dell'utente, analizzando i dati e individuando le correzioni da apportare, escludendo i campi problematici o probabilmente inutili, derivando nuovi attributi se necessario e migliorando le prestazioni attraverso tecniche di screening intelligenti. È possibile utilizzare l'algoritmo in modo completamente **automatico**, consentendo all'algoritmo di scegliere ed applicare le correzioni oppure è possibile utilizzare la modalità **interattiva**, in cui viene visualizzata un'anteprima delle modifiche prima che vengano apportate, in modo che sia possibile accettarle o rifiutarle in base alle proprie esigenze.

L'utilizzo di ADP consente di predisporre i dati per la creazione del modello in modo semplice e rapido, senza che sia necessario conoscere i concetti statistici impiegati. La creazione e il calcolo del punteggio dei modelli tenderanno a essere più rapidi; inoltre, l'utilizzo di ADP migliora la robustezza dei processi di modellazione automatica.

Nota: quando ADP prepara un campo per l'analisi, crea un nuovo campo contenente le regolazioni o le trasformazioni, invece di sostituire i valori e le proprietà esistenti dal campo precedente. Il vecchio campo non viene utilizzato in ulteriori analisi; il suo ruolo è impostato su Nessuno. Nota inoltre che qualsiasi informazione di valore mancante dall'utente non viene trasferita a questi campi di nuova creazione e gli eventuali valori mancanti nel nuovo campo sono mancanti di sistema.

Esempio. Una compagnia di assicurazioni con poche risorse per indagare sulle richieste di indennizzo dei proprietari immobiliari vuole creare un modello per evidenziare le richieste sospette e potenzialmente fraudolente. Prima di procedere, viene effettuata la preparazione automatica dei dati per la creazione del modello. Dal momento che la compagnia ha necessità di esaminare le trasformazioni proposte prima che queste vengano applicate, utilizzerà la preparazione automatica dati in modalità interattiva.

Un gruppo industriale automobilistico tiene traccia delle vendite per un'ampia gamma di autoveicoli personali. Nel tentativo di identificare modelli a basso e alto rendimento è possibile stabilire una relazione tra la vendita dei veicoli e le rispettive caratteristiche. Verrà utilizzata la preparazione automatica dei dati per l'analisi e verranno creati modelli utilizzando i dati "prima" e "dopo" la preparazione per scoprire come cambiano i risultati.

Qual è il proprio l'obiettivo? La Preparazione automatica dati consiglia una serie di passi di preparazione dei dati che influiscono sulla velocità con cui altri algoritmi creano modelli e ne migliorano il potere predittivo. può comprendere la trasformazione, la creazione e la selezione delle funzioni. Anche l'obiettivo può essere trasformato. È possibile specificare le priorità di creazione dei modelli su cui deve concentrarsi il processo di preparazione dei dati.

- **Bilancia velocità e accuratezza.** Questa opzione prepara i dati in modo da dare la stessa priorità alla velocità di elaborazione dei dati da parte degli algoritmi di creazione del modello e alla precisione delle previsioni.
- **Ottimizza per velocità.** Questa opzione prepara i dati in modo da dare la priorità alla velocità di elaborazione dei dati da parte degli algoritmi di creazione del modello. Selezionare questa opzione quando si utilizzano insiemi di dati molto grandi o quando si desidera ottenere una risposta rapida.
- **Ottimizza per precisione.** Questa opzione prepara i dati in modo da dare la priorità alla precisione delle previsioni generate dagli algoritmi di creazione del modello.
- **Analisi personalizzata.** Selezionare questa opzione se si desidera modificare manualmente l'algoritmo nella scheda Impostazioni. Si noti che questa impostazione viene selezionata automaticamente se in seguito si apportano modifiche incompatibili con uno degli altri obiettivi alle opzioni della scheda Impostazioni.

Per Ottenere La Preparazione Automatica Dei Dati

Dai menu, scegliere:

1. Dai menu, scegliere:

Trasforma > **Preparare i dati per la modellazione** > **Automatico ...**

2. Fare clic su **Esegui**.

Facoltativamente, è possibile:

- Specificare un obiettivo nella scheda Obiettivi.
- Specificare le assegnazioni di campo nella scheda Campi.
- Specificare delle impostazioni avanzate nella scheda Impostazioni.

Per Ottenere La Preparazione Dei Dati Interattivi

1. Dai menu, scegliere:

   Trasforma > **Preparare i dati per la modellazione** > **Interactive ...**

2. Clicca su **Analizza** nella barra degli strumenti nella parte superiore della finestra di dialogo.
3. Fare clic sulla scheda Analisi e rivedere i passi di preparazione dei dati suggeriti.
4. Se soddisfatto, fare clic su **Esegui**. In caso contrario, fare clic su **Cancella analisi**, modificare le impostazioni desiderate e fare clic su **Analizza**.

Facoltativamente, è possibile:

- Specificare un obiettivo nella scheda Obiettivi.
- Specificare le assegnazioni di campo nella scheda Campi.
- Specificare delle impostazioni avanzate nella scheda Impostazioni.
- Salvare i passi di preparazione dei dati suggeriti in un file XML facendo clic su **Salva XML**.

Scheda Campi

La scheda Campi indica i campi che è necessario preparare per eseguire ulteriori analisi.

Utilizza ruoli predefiniti. Questa opzione utilizza le informazioni contenute nei campi esistenti. Se esiste un solo campo con un ruolo come Obiettivo, sarà utilizzato come obiettivo; altrimenti non vi sarà alcun obiettivo. Tutti i campi con un ruolo predefinito come Input saranno utilizzati come input. È necessario almeno un campo di input.

Utilizza assegnazioni campi personalizzate. Quando si ignorano i ruoli dei campi spostando i campi dai relativi elenchi predefiniti, la finestra di dialogo visualizza automaticamente questa opzione. Quando si effettuano delle assegnazioni campi personalizzate, specificare i campi seguenti:

- **Obiettivo (facoltativo).** Se i modelli da creare richiedono un obiettivo, selezionare il campo obiettivo. Questa operazione equivale a impostare il ruolo del campo su Obiettivo.
- **Input.** Selezionare uno o più campi di input. Questa operazione equivale a impostare il ruolo del campo su Input.

Scheda Impostazioni

La scheda Impostazioni contiene vari gruppi di impostazioni differenti che è possibile modificare per definire il modo in cui l'algoritmo elabora i dati. Se si apportano modifiche alle impostazioni predefinite che risultano incompatibili con gli altri obiettivi, la scheda Obiettivo viene aggiornata automaticamente per selezionare l'opzione **Personalizza analisi**.

Prepara Date & Times

Molti algoritmi di modellazione non sono in grado di gestire direttamente i dettagli relativi a date e ore; queste impostazioni consentono di derivare nuovi dati sulle durate utilizzabili come input per i modelli dalle date e dalle ore indicate nei dati esistenti. I campi contenenti date e ore devono essere predefiniti con tipi di archiviazione data o ora. L'uso dei campi data e ora originali come input per i modelli in seguito alla preparazione automatica dei dati non è consigliato.

Prepara date e ore per la modellazione. Se si deseleziona questa opzione vengono disabilitati tutti gli altri comandi Prepara date e ore ma vengono mantenute le selezioni.

Calcola tempo trascorso fino alla data di riferimento. Produce il numero di anni/mesi/giorni a partire da una data di riferimento per ciascuna variabile che contiene delle date.

- **Data di riferimento.** Specifica la data a partire da cui sarà calcolata la durata relativamente alle informazioni sulla data presenti nei dati di input. Selezionando **Data odierna**, viene sempre utilizzata la data di sistema corrente quando viene eseguito ADP. Per utilizzare una data specifica, selezionare **Data fissa** e immettere la data desiderata.
- **Unità per la durata della data.** Specificare se ADP deve decidere automaticamente l'unità per la durata della data oppure selezionarne una da **Unità fisse** in Anni, mesi o Giorni.

Calcola tempo trascorso fino all'ora di riferimento. Produce il numero di ore/minuti/secondi a partire da un'ora di riferimento per ciascuna variabile che contiene delle ore.

- **Ora di riferimento.** Specifica l'ora a partire dalla quale sarà calcolata la durata relativamente alle informazioni sull'ora presenti nei dati di input. Se si seleziona **Ora corrente**, viene sempre utilizzata l'ora corrente del sistema per l'esecuzione di ADP. Per utilizzare un'ora specifica, selezionare **Ora fissa** e immettere l'ora desiderata.
- **Unità per la durata dell'ora.** Specificare se ADP deve decidere automaticamente l'unità per la durata dell'ora oppure selezionarne una da **Unità fisse** in Ore, minuti o Secondi.

Estrai elementi di tempo ciclico. Utilizzare queste impostazioni per suddividere un singolo campo data o ora in uno o più campi. Ad esempio, se vengono selezionate tutte e tre le caselle di controllo della data, il campo della data di input "1954-05-23" viene suddiviso in tre campi: 1954, 5 e 23, ciascuno dei quali utilizza il suffisso definito nel pannello **Nomi di campi** ed il campo della data originale viene ignorato.

- **Estrai dalle date.** Per qualsiasi input di data, specificare se si desidera estrarre gli anni, i mesi, i giorni o una combinazione dei tre elementi.
- **Estrai dalle ore.** Per qualsiasi input di ora, specificare se si desidera estrarre le ore, i minuti, i secondi o una combinazione dei tre elementi.

Escludi campi

La scarsa qualità dei dati può influire sulla precisione delle previsioni; pertanto, è possibile specificare il livello di qualità accettabile per le funzioni di input. Tutti i campi che sono costanti o hanno il 100% dei valori mancanti vengono esclusi automaticamente.

Escludi campi di input a bassa qualità. Se si deseleziona questa opzione vengono disabilitati tutti gli altri comandi Escludi campi ma vengono mantenute le selezioni.

Escludi campi con troppi valori mancanti. I campi con una percentuale di valori mancanti superiore a quella specificata vengono eliminati dalla successiva analisi. Specificare un valore superiore o uguale a 0, che equivale a deselezionare questa opzione, e inferiore o uguale a 100, benché i campi con tutti i valori mancanti vengano automaticamente esclusi. Il valore di default è 50.

Escludi campi nominali con troppe categorie univoche. I campi nominali con un numero di categorie superiore a quello specificato vengono eliminati dalla successiva analisi. Specificare un intero positivo. Il valore predefinito è 100. È utile per rimuovere automaticamente i campi che contengono informazioni univoche del record provenienti dalla modellazione, come ID, indirizzo o nome.

Escludi campi categoriali con troppi valori in una categoria singola. I campi ordinali e nominali con una categoria contenente una percentuale di record superiore a quella specificata vengono eliminati dalla successiva analisi. Specificare un valore superiore o uguale a 0, che equivale a deselezionare questa opzione, e inferiore o uguale a 100, benché i campi costanti vengano automaticamente esclusi. Il valore predefinito è 95.

Regolare Misurazione

Adatta livello di misurazione. La deselezione di questa opzione disabilita tutti gli altri controlli Regolare Misurazione mantenendo le selezioni.

Livello di misurazione. Specificare se il livello di misurazione dei campi continui con valori "troppo pochi" può essere regolato su ordinali, e i campi ordinali con valori "troppi" possono essere regolati in continuo.

- **Numero massimo di valori per i campi ordinali.** I campi ordinali con più del numero specificato di categorie sono recintati come campi continui. Specificare un intero positivo. Il valore predefinito è 10. Questo valore deve essere maggiore o uguale al numero minimo di valori per i campi continui.
- **Numero minimo di valori per i campi continui.** I campi continui con meno del numero specificato di valori univoci sono recintati come campi ordinali. Specificare un intero positivo. Il valore predefinito è 5. Questo valore deve essere inferiore o uguale al numero massimo di valori per i campi ordinali.

Migliorare la qualità dei dati

Preparare i campi per migliorare la qualità dei dati. La deselezione di questa opzione disabilita tutte le altre Migliori controlli di qualità dei dati mantenendo le selezioni.

Gestione dei valori anomali. Specificare se sostituire gli outliers per gli input e target; in caso affermativo, specificare un criterio di cutoff dell'outlier, misurato nelle deviazioni standard, e un metodo per la sostituzione degli outliers. Gli outliers possono essere sostituiti da trimming (impostazione al valore di cutoff) oppure impostandoli come valori mancanti. Gli eventuali outlier impostati sui valori mancanti seguono le impostazioni di gestione del valore mancanti selezionate di seguito.

Sostituire Valori mancanti. Specificare se sostituire i valori mancanti di campi continui, nominali o ordinali.

Riordina campi nominali. Selezioniamo questo per ricuperare i valori dei campi nominali (set) da quelli più piccoli (meno frequentemente) alla categoria più grande (più frequentemente presente). I nuovi valori di campo iniziano con 0 come categoria meno frequente. Da notare che il nuovo campo sarà numerico anche se il campo originale è una stringa. Ad esempio, se i valori dei dati di un campo nominale sono "A", "A", "A", "B", "C", "C", la preparazione automatica dati ricodifica "B" in 0, "C" in 1 e "A" in 2.

Campi di Rescale

Campi di scala. La deselezione di questa opzione disabilita tutti gli altri controlli Campi Rescale mantenendo le selezioni.

Peso analisi. Questa variabile contiene pesi di analisi (regressione o campionamento). I pesi di analisi vengono utilizzati per rendere conto delle differenze di varianza tra i livelli del campo di destinazione. Seleziona un campo continuo.

Campi di input continui. In questo modo si normalizzerà i campi di input continui utilizzando una trasformazione **z - score** o **min / max**. Gli input di Rescaling sono utili soprattutto quando si seleziona **Perform feature construction** sulle impostazioni di Seleziona e Costrutto.

- **Trasformazione punteggio Z.** Utilizzando la media osservata e la deviazione standard come stime dei parametri di popolazione, i campi sono standardizzati e poi i punteggi *z* sono mappati ai valori corrispondenti di una distribuzione normale con la **media finale** e **deviazione standard finale**. Specificare un numero per **Finale finale** e un numero positivo per **deviazione standard finale**. Le impostazioni predefinite sono rispettivamente 0 e 1, corrispondenti a ridimensionamento standardizzato.
- **Trasformazione Min/Max.** Utilizzando le stime minime e massime osservate come stime dei parametri di popolazione, i campi vengono mappati ai valori corrispondenti di una distribuzione uniforme con il **minimo** e **Massimo**. Specificare i numeri con **Massimo** maggiore di **Minimo**.

Target continuo. Questo trasforma un obiettivo continuo utilizzando la trasformazione Box - Cox in un campo che ha una distribuzione approssimativamente normale con la **media finale** e **deviazione standard finale**. Specificare un numero per **Finale finale** e un numero positivo per **deviazione standard finale**. I default sono rispettivamente 0 e 1.

Nota: se un obiettivo è stato trasformato da ADP, i modelli successivi costruiti utilizzando l'obiettivo trasformato segnano le unità trasformate. Per interpretare e utilizzare questi risultati è necessario riconvertire il valore previsto nella scala originale. Vedi l'argomento per ulteriori informazioni. Vedi l'argomento “Trasforma all'indietro i punteggi” a pagina 18 per ulteriori informazioni.

Trasforma i campi

Per migliorare la potenza predittiva dei tuoi dati è possibile trasformare i campi di input.

Campo di trasformazione per la modellazione. La deselezione di questa opzione disabilita tutti gli altri controlli dei Campi Transform mantenendo le selezioni.

Campi di input categoriali Le seguenti opzioni sono disponibili:

- **Unisci categorie sparse per aumentare al massimo l'associazione all'obiettivo.** Selezionare questa opzione per creare un modello più gestibile riducendo il numero dei campi da elaborare in associazione all'obiettivo. Le categorie simili vengono identificate in base alla relazione tra input e obiettivo. Le categorie che non sono significativamente diverse (cioè avere un valore *p*superiore al valore specificato) sono fuse. Specificare un valore superiore a 0 e inferiore o uguale a 1. Se tutte le categorie sono fuse in una, le versioni originali e derivate del campo sono escluse da ulteriori analisi perché non hanno alcun valore come predittore.
- **Quando non ci sono obiettivi, unisci le categorie sparse in base ai conteggi.** Se il dataset non ha alcun obiettivo, è possibile scegliere di unire categorie sparse di campi ordinali e nominali. Il metodo di uguale frequenza viene utilizzato per unire categorie con meno della percentuale minima specificata del numero totale di record. Specificare un valore maggiore o uguale a 0 e minore o uguale a 100. Il valore predefinito è 10. La fusione si arresta quando non ci sono categorie con meno del minimo specificato di casi, o quando ci sono solo due categorie rimaste.

Campi di input continui. Se il dataset include un target categoriale, è possibile bistare input continui con associazioni forti per migliorare le prestazioni di elaborazione. I bidoni vengono creati in base alle proprietà di "sottoinsiemi omogenei", che vengono identificati dal metodo Scheffe utilizzando il valore specificato *p*come alfa per il valore critico per la determinazione dei sottoinsiemi omogenei. Specificare un valore superiore a 0 e inferiore o uguale a 1. Il valore predefinito è 0.05. Se con l'operazione di categorizzazione si ottiene un unico intervallo per un determinato campo, la versione originale e categorizzata del campo vengono escluse perché sono prive di valore come predittori.

Nota: il Binning in ADP differisce dal binning ottimale. Il binning ottimale utilizza informazioni entropiche per convertire un campo continuo in un campo categoriale; questo deve ordinare i dati e memorizzarlo tutto in memoria. ADP utilizza sottoinsiemi omogenei per bin un campo continuo, il che significa che il binning ADP non ha bisogno di ordinare i dati e non memorizza tutti i dati in memoria. Quando si utilizza il metodo dei sottoinsiemi omogenei per categorizzare un campo continuo, il numero di categorie dopo la categorizzazione è sempre inferiore o uguale al numero di categorie dell'obiettivo.

Selezionare e Costruire

Per migliorare la potenza predittiva dei tuoi dati, puoi costruire nuovi campi in base ai campi esistenti.

Effettua selezione delle funzioni. Un input continuo viene rimosso dall'analisi se il valore *p*per la sua correlazione con l'obiettivo è maggiore del valore specificato *p*.

Esegui creazione funzioni. Selezionare questa opzione per ricavare nuove funzioni da una combinazione di diverse funzioni esistenti. Le vecchie funzioni non vengono utilizzate in ulteriori analisi. Questa opzione viene applicata solo alle funzioni di input continue quando l'obiettivo è continuo oppure non esiste.

Nomi di campi

Per individuare facilmente le funzioni nuove e trasformate, ADP crea e applica nuovi nomi, prefissi o suffissi di base. I nomi si possono modificare in modo da renderli più pertinenti rispetto alle esigenze e ai dati dell'utente.

Campi trasformati e creati. Specificare le estensioni dei nomi da applicare ai campi obiettivo e di input trasformati.

Specificare inoltre il nome del prefisso da applicare a tutte le funzioni da creare con le impostazioni Seleziona e Crea. Il nuovo nome viene creato apponendo un suffisso numerico al nome radice del prefisso. Il formato del numero dipende dal numero di nuove funzioni da derivare, ad esempio:

- le funzioni create da 1 a 9 verranno denominate: da feature1 a feature9.

- le funzioni create da 10 a 99 verranno denominate: da feature01 a feature99.
- le funzioni create da 100 a 999 verranno denominate: da feature001 a feature999 e così via.

In questo modo, le funzioni create saranno organizzate in un ordine logico indipendentemente dal loro numero.

Durate calcolate da date e ore. Specificare le estensioni dei nomi da applicare alle durate calcolate a partire da date e ore.

Elementi ciclici estratti da date e ore. Specificare le estensioni dei nomi da applicare agli elementi ciclici estratti da date e ore.

Applicazione e Saving Trasformazioni

A seconda che si stia utilizzando i dialoghi Interactive o Automatic Data Preparazione, le impostazioni per applicare e salvare le trasformazioni sono leggermente diverse.

Impostazioni di preparazione dei dati interattivi Applicazioni

Dati trasformati. Queste impostazioni specificano dove salvare i dati trasformati.

- **Aggiungere nuovi campi al dataset attivo.** Qualsiasi campo creato dalla preparazione dei dati automatizzati viene aggiunto come nuovi campi al dataset attivo. **Aggiorna i ruoli per i campi analizzati** imposterà il ruolo a Nessuno per eventuali campi esclusi da ulteriori analisi mediante la preparazione dei dati automatizzati.
- **Creare un nuovo dataset o un file contenente i dati trasformati.** I campi consigliati dalla preparazione dei dati automatizzati vengono aggiunti ad un nuovo dataset o file. **Includi campi non analizzati** aggiunge i campi nel dataset originale che non sono stati specificati nella scheda Campi al nuovo dataset. Questo è utile per trasferire i campi contenenti informazioni non utilizzate nella modellazione, come ID o indirizzo, o nome, nel nuovo dataset.

Impostazioni automatiche di preparazione dei dati e impostazioni di salvataggio

Il Gruppo Dati Trasformati è lo stesso di Interactive Data Preparazione. Nella preparazione automatica dei dati sono disponibili le seguenti opzioni aggiuntive:

Applicare le trasformazioni. Nei dialoghi automatici di preparazione dei dati, la deselezione di questa opzione disabilita tutti gli altri controlli Apply e Salva mantenendo le selezioni.

Salva trasformazioni come sintassi. Questo salva le trasformazioni consigliate come sintassi del comando ad un file esterno. Dialogo Interactive Data Preparazione non ha questo controllo perché incollare le trasformazioni come sintassi del comando alla finestra di sintassi se si fa clic su **Paste**.

Salva trasformazioni come XML. Questo salva le trasformazioni consigliate come XML in un file esterno, che può essere unito al modello PMML utilizzando `TMS MERGE` o applicato ad un altro dataset utilizzando `TMS IMPORT`. Dialogo Interactive Data Preparazione non ha questo controllo perché salverà le trasformazioni come XML se si fa clic su **Salva XML** nella barra degli strumenti nella parte superiore della finestra di dialogo.

Scheda Analisi

Nota: la scheda Analisi viene utilizzata nella finestra di dialogo Preparazione interattiva dati per consentire di esaminare le trasformazioni consigliate. La finestra di dialogo Preparazione automatica dati non prevede questo passaggio.

1. Quando le impostazioni di ADP sono soddisfacenti (comprese le eventuali modifiche apportate alla schede Obiettivo, Campi e Impostazioni), fare clic su **Analizza dati**; l'algoritmo applica le impostazioni ai dati immessi e visualizza i risultati nella scheda Analisi.

La scheda Analisi contiene output in formato tabellare e grafico che riassume l'elaborazione dei dati e visualizza le raccomandazioni su eventuali modifiche o miglioramenti da apportare ai dati per il calcolo del punteggio. Questo consente di esaminare e accettare o rifiutare tali raccomandazioni.

La scheda Analisi è composta da due riquadri, la visualizzazione principale a sinistra e quella collegata o ausiliaria a destra. Le visualizzazioni principali sono tre:

- Riepilogo elaborazione campi (impostazione predefinita). Per ulteriori informazioni, consultare l'argomento “Riepilogo elaborazione campi ” a pagina 13.
- Campi. Per ulteriori informazioni, consultare l'argomento “Campi ” a pagina 14.
- Riepilogo delle azioni Per ulteriori informazioni, consultare l'argomento “Riepilogo delle azioni ” a pagina 15.

Le viste collegate/ausiliarie sono quattro:

- Potere predittivo (impostazione predefinita). Per ulteriori informazioni, consultare l'argomento “Potere predittivo ” a pagina 15.
- Tabella campi. Per ulteriori informazioni, consultare l'argomento “Tabella Campi ” a pagina 15.
- Dettagli campo. Per ulteriori informazioni, consultare l'argomento “Dettagli campo ” a pagina 16.
- Dettagli dell'azione. Per ulteriori informazioni, consultare l'argomento “Dettagli azione ” a pagina 17.

Collegamenti tra le visualizzazioni

All'interno della visualizzazione principale, il testo sottolineato nelle tabelle controlla la visualizzazione nella visualizzazione collegata. Se si fa clic sul testo è possibile visualizzare i dettagli di un determinato campo, insieme di campi o fase di elaborazione. Il collegamento selezionato per ultimo è visualizzato con un colore più scuro per facilitare l'individuazione del collegamento tra il contenuto dei due riquadri.

Reimpostazione delle visualizzazioni

Per visualizzare nuovamente le raccomandazioni originali della scheda Analisi e annullare le eventuali modifiche apportate alle visualizzazioni Analisi, fare clic su **Reimposta** nella parte inferiore del riquadro di visualizzazione principale.

Riepilogo elaborazione campi

La tabella Riepilogo elaborazione campi offre una snapshot dell'impatto complessivo stimato dell'elaborazione, comprese le modifiche dello stato delle funzioni e il numero di funzioni create.

Si noti che non viene effettivamente creato alcun modello e, pertanto, non vi è alcuna misura o grafico della variazione del potere predittivo generale prima e dopo la preparazione dei dati; è possibile invece visualizzare i grafici del potere predittivo dei singoli predittori consigliati.

La tabella riporta le seguenti informazioni:

- Il numero di campi obiettivo.
- Il numero di predittori (input) originali.
- I predittori consigliati per l'uso nell'analisi e nella modellazione. Sono inclusi il numero totale di campi consigliati; il numero di campi originali, non trasformati consigliati; il numero di campi trasformati consigliati (escluse le versioni intermedie di qualsiasi campo, i campi derivati dai predittori di data/ora e i predittori creati); il numero di campi consigliati derivati dai campi data/ora e il numero di predittori creati consigliati.
- Il numero di predittori di input non consigliato per l'uso in nessuna forma, che si tratti della forma originale, come campo derivato o come input per un predittore creato.

Se le informazioni dei **Campi** sono sottolineate, farvi clic sopra per visualizzare ulteriori dettagli in una visualizzazione collegata. I dettagli relativi a **Obiettivo**, **Funzioni di input** e **Funzioni di input non utilizzate** sono riportati nella visualizzazione collegata Tabella campi. Per ulteriori informazioni, consultare l'argomento “Tabella Campi ” a pagina 15. **Funzioni consigliate per l'utilizzo in analisi** vengono visualizzate nella vista Predictive Power linked. Per ulteriori informazioni, consultare l'argomento “Potere predittivo ” a pagina 15.

Campi

La visualizzazione principale Campi mostra i campi elaborati e indica se ADP ne consiglia o meno l'utilizzo nei modelli a valle. È possibile ignorare le raccomandazioni di tutti i campi, ad esempio per escludere funzioni create o includere funzioni di cui ADP consiglia l'esclusione. Se un campo è stato trasformato, è possibile decidere se accettare la trasformazione suggerita o utilizzare la versione originale.

La visualizzazione Campi è composta da due tabelle, una per l'obiettivo e una per i predittori elaborati o creati.

Tabella Obiettivo

La tabella **Obiettivo** è visualizzata solo se nei dati è stato definito un obiettivo.

La tabella contiene due colonne:

- **Nome.** Si tratta del nome o dell'etichetta del campo obiettivo. Viene sempre utilizzato il nome originale, anche se il campo è stato trasformato.
- **Livello di misurazione.** In questa colonna è visualizzata l'icona che rappresenta il livello di misurazione; passare il puntatore del mouse sopra l'icona per visualizzare un'etichetta (continuo, ordinale, nominale e così via) che descrive i dati.

  Se l'obiettivo è stato trasformato, la colonna **Livello di misurazione** riflette la versione trasformata finale. *Nota*: non è possibile disattivare le trasformazioni per l'obiettivo.

Tabella Predittori

La tabella **Predittori** è sempre visualizzata. Ogni riga della tabella rappresenta un campo. Per impostazione predefinita, le righe sono ordinate in modo decrescente in base al potere predittivo.

Per le funzioni ordinarie, il nome originale viene sempre utilizzato come nome della riga. Nella tabella sono riportate sia le versioni originali che quelle derivate dei campi data/ora (in righe separate); la tabella comprende anche i predittori creati.

Si noti che le versioni trasformate dei campi visualizzate nella tabella rappresentano sempre le versioni finali.

Per impostazione predefinita, nella tabella Predittori sono visualizzati solo i campi consigliati. Per visualizzare gli altri campi, selezionare la casella **Includi campi non raccomandati nella tabella** sopra la tabella; in questo modo, tali campi saranno visualizzati in fondo alla tabella.

La tabella contiene le seguenti colonne:

- **Versione da usare.** Questa colonna visualizza un elenco a discesa che controlla se un campo verrà utilizzato a valle e se usare le trasformazioni suggerite. Per impostazione predefinita, l'elenco rispecchia le raccomandazioni.

  Per i predittori ordinari trasformati, l'elenco a discesa contiene tre opzioni: **Trasformata**, **Originale** e **Non utilizzare**.

  Per i predittori ordinari non trasformati, le scelte sono: **Originale** e **Non utilizzare**.

  Per i campi data/ora derivati e per i predittori creati, le scelte sono: **Trasformata** e **Non utilizzare**.

  Per i campi data originali l'elenco a discesa è disattivato e impostato su **Non utilizzare**.

  Nota: per i predittori con versioni originale e trasformata, il passaggio dalla versione **Originale** a quella **Trasformata** aggiorna automaticamente le impostazioni **Livello di misurazione** e **Potere predittivo** per tali funzioni.
- **Nome.** Ogni nome di campo è un collegamento. Fare clic su un nome per visualizzare ulteriori informazioni relative al campo nella visualizzazione collegata. Per ulteriori informazioni, consultare l'argomento “Dettagli campo ” a pagina 16.
- **Livello di misurazione.** In questa colonna è visualizzata l'icona che rappresenta il tipo di dati; passare il puntatore del mouse sopra l'icona per visualizzare un'etichetta (continuo, ordinale, nominale e così via) che descrive i dati.

- **Potere predittivo.** Il potere predittivo è visualizzato solo per i campi consigliati da ADP. Questa colonna non è visualizzata se non è stato definito un obiettivo. Il potere predittivo è compreso tra 0 e 1, e i valori più elevati indicano predittori "migliori". In generale, il potere predittivo è utile per confrontare i predittori all'interno di un'analisi ADP, ma non deve essere effettuato alcun confronto tra i valori del potere predittivo di analisi diverse.

Riepilogo delle azioni

Per ogni azione svolta dalla preparazione automatica dati, i predittori di input vengono trasformati e/o eliminati tramite filtri; i campi che sopravvivono a un'azione vengono utilizzati nella successiva. I campi che superano tutti i passaggi sono quelli di cui si consiglia l'utilizzo nella modellazione, mentre gli input a predittori trasformati e creati vengono esclusi.

Il Riepilogo delle azioni è una semplice tabella che elenca le azioni di elaborazione svolte da ADP. Se vi è un'**Azione** sottolineata, farvi clic sopra per visualizzare ulteriori dettagli sulle azioni intraprese in una visualizzazione collegata. Per ulteriori informazioni, consultare l'argomento “Dettagli azione ” a pagina 17.

Nota: sono visualizzate solo le versioni originale e trasformata definitiva di ciascun campo, non le versioni intermedie utilizzate durante l'analisi.

Potere predittivo

Visualizzato per impostazione predefinita la prima volta che viene eseguita l'analisi o quando si seleziona **Predittori consigliati per l'uso nell'analisi** nella visualizzazione principale Riepilogo elaborazione campi, il grafico mostra il potere predittivo dei predittori consigliati. I campi sono ordinati in base al potere predittivo, a partire dal campo con il valore più elevato.

Per le versioni trasformate dei predittori ordinari, il nome del campo riflette il suffisso scelto nel pannello Nomi di campi della scheda Impostazioni; ad esempio: *_transformed*.

Dopo i singoli nomi dei campi sono visualizzate le icone del livello di misurazione.

Il potere predittivo di ciascun predittore consigliato viene calcolato da una regressione lineare o modello naïve Bayes, in base al fatto che l'obiettivo sia continuo o relativo alla categoria.

Tabella Campi

Visualizzata quando si fa clic su **Obiettivo**, **Predittori** o **Predittori non utilizzati** nella visualizzazione principale Riepilogo elaborazione campi, la visualizzazione Tabella campi mostra una semplice tabella con un elenco delle funzioni pertinenti.

La tabella contiene due colonne:

- **Nome.** Il nome del predittore.

  Per gli obiettivi viene utilizzato il nome o l'etichetta originale del campo, anche se l'obiettivo è stato trasformato.

  Per le versioni trasformate dei predittori ordinari, il nome riflette il suffisso scelto nel pannello Nomi di campi della scheda Impostazioni; ad esempio: *_transformed*.

  Per i campi derivati da date ed ore, viene utilizzato il nome della versione trasformata finale; ad esempio: *bdate_years*.

  Per i predittori creati, viene utilizzato il nome del predittore creato; ad esempio: *Predictor1*.
- **Livello di misurazione.** Visualizza l'icona che rappresenta il tipo di dati.

  Per l'Obiettivo, il **Livello di misurazione** rispecchia sempre la versione trasformata se l'obiettivo è stato trasformato, ad esempio passando da ordinale (insieme ordinato) a continuo (intervallo, scala) o viceversa.

Dettagli campo

Visualizzata quando si fa clic su un **Nome** nella visualizzazione principale Campi, la visualizzazione Dettagli campo contiene la distribuzione, i valori mancanti e gli eventuali grafici del potere predittivo per il campo selezionato. Inoltre, vengono visualizzati anche la cronologia di elaborazione per il campo ed il nome del campo trasformato (se applicabile).

Per ogni grafico impostato sono visualizzate due versioni affiancate per confrontare il campo con e senza l'applicazione delle trasformazioni; se non esiste una versione trasformata del campo, il grafico viene visualizzato solo per la versione originale. Per i campi data o ora derivati e i predittori creati, i grafici sono visualizzati solo per il nuovo predittore.

Nota: se un campo viene escluso perché dispone di troppe categorie, viene visualizzata solo la cronologia di elaborazione.

Grafico della distribuzione

La distribuzione dei campi continui è rappresentata sotto forma di istogramma a cui è sovrapposta una curva normale e con una linea verticale di riferimento per il valore medio; i campi categoriali sono visualizzati sotto forma di grafico a barre.

Gli istogrammi sono dotati di etichette che mostrano la deviazione standard e l'asimmetria; tuttavia, l'asimmetria non è visualizzata se il numero massimo dei valori è 2 o se la varianza del campo originale è inferiore a 10-20.

Passare il puntatore del mouse sul grafico per visualizzare la media degli istogrammi o il numero e la percentuale sul totale dei record per le per le categorie dei grafici a barre.

Grafico dei valori mancanti

I grafici a torta confrontano la percentuale dei valori mancanti con e senza l'applicazione delle trasformazioni; le etichette del grafico mostrano la percentuale.

Se ADP ha utilizzato la gestione dei valori mancanti, il grafico a torta dopo la trasformazione comprende anche il valore di sostituzione (cioè il valore utilizzato al posto di quelli mancanti) sotto forma di etichetta.

Passare il puntatore del mouse sopra il grafico per visualizzare il numero dei valori mancanti e la percentuale del numero totale di record.

Grafico del potere predittivo

Per i campi consigliati, i grafici a barre mostrano il potere predittivo prima e dopo la trasformazione. Se l'obiettivo è stato trasformato, il potere predittivo calcolato è relativo all'obiettivo trasformato.

Nota: i grafici del potere predittivo non vengono visualizzati se non è definito alcun obiettivo oppure se si fa clic sull'obiettivo nel pannello della visualizzazione principale.

Passare il puntatore del mouse sul grafico per visualizzare il valore del potere predittivo.

Tabella Cronologia elaborazione

La tabella mostra come è stata derivata la versione trasformata di un campo. Le azioni intraprese da ADP sono elencate nell'ordine in cui sono state eseguite; tuttavia, per alcuni passaggi è possibile che siano state intraprese più azioni per un determinato campo.

Nota: questa tabella non viene visualizzata per i campi che non sono stati trasformati.

Le informazioni della tabella sono suddivise in due o tre colonne:

- **Azione.** Il nome dell'azione. Ad esempio, Predittori continui. Per ulteriori informazioni, consultare l'argomento “Dettagli azione ” a pagina 17.
- **Dettagli.** L'elenco delle procedure eseguite. Ad esempio, Trasforma in unità standard.
- **Funzione.** Viene visualizzata solo per i predittori creati e visualizza la combinazione lineare dei campi di input, ad esempio .06*age + 1.21*height.

Dettagli azione

Visualizzata quando si seleziona un'**Azione** sottolineata nella visualizzazione principale Riepilogo delle azioni, la visualizzazione collegata Dettagli azione mostra informazioni generali e specifiche di un'azione per ogni fase di elaborazione effettuata; i dettagli relativi a un'azione specifica sono visualizzati per primi.

Per ogni azione, la descrizione viene utilizzata come titolo nella parte superiore della visualizzazione collegata. I dettagli specifici delle singole azioni sono visualizzati sotto al titolo e possono comprendere il numero di predittori derivati, i campi riformulati, le trasformazioni dell'obiettivo, le categorie unite o riordinate e i predittori creati o esclusi.

A ogni azione, il numero di predittori utilizzato nell'elaborazione può variare, ad esempio a causa dell'esclusione o dell'unione di predittori.

Nota: se un'azione è stata disattivata o se non è stato specificato alcun obiettivo, quando si fa clic sull'azione nella visualizzazione principale Riepilogo delle azioni, viene visualizzato un messaggio di errore invece dei dettagli dell'azione.

Le possibili azioni sono nove, ma non tutte sono necessariamente attive per ogni analisi.

Tabella Campi testo

La tabella mostra il numero di:

- Predittori esclusi dall'analisi.

Tabella Predittori data e ora

La tabella mostra il numero di:

- Durate derivate da predittori di data e ora.
- Elementi di data e ora.
- Predittori di data e ora derivati in totale.

La data o l'ora di riferimento è visualizzata come nota a piè di pagina se sono state calcolate delle durate delle date.

Tabella Screening dei predittori

La tabella mostra il numero dei seguenti predittori esclusi dall'elaborazione:

- Costanti.
- Predittori con troppi valori mancanti.
- Predittori con troppi casi in un'unica categoria.
- Campi nominali (insiemi) con troppe categorie.
- Predittori esclusi in totale.

Tabella Controlla livello di misurazione

La tabella mostra il numero dei campi riformulati, suddivisi in:

- Campi ordinali (insiemi ordinati) riformulati come campi continui.
- Campi continui riformulati come campi ordinali.
- Numero totale riformulato.

Se nessun campo di input (obiettivo o predittore) è continuo o ordinale, il totale viene visualizzato come un piè di pagina.

Tabella Valori anomali

La tabella mostra il conteggio delle modalità con cui sono stati gestiti i valori anomali.

- Il numero di campi continui per cui i valori anomali sono stati rilevati e tagliati oppure il numero di campi continui per cui i valori anomali sono stati rilevati ed impostati su mancanti, in base alle impostazioni nel pannello Prepara input e obiettivo nella scheda Impostazioni.

- Il numero dei campi continui esclusi perché costanti dopo la gestione dei valori anomali.

Un piè di pagina indica il punto di interruzione dei valori anomali, mentre viene mostrato un altro piè di pagina se nessun campo di input (obiettivo o predittore) è continuo.

Tabella Valori mancanti

La tabella mostra il numero dei campi i cui valori mancanti sono stati sostituiti, suddivisi in:

- Obiettivo. Se non viene specificato alcun obiettivo, questa riga non è visualizzata.
- Predittori. A sua volta suddiviso nel numero di nominali (insieme), ordinali (insieme ordinato) e continui.
- Il numero totale di valori mancanti sostituiti.

Tabella Obiettivo

La tabella indica se l'obiettivo è stato trasformato, illustrato come:

- Trasformazione di Box-Cox alla normalità. Questo valore è ulteriormente suddiviso in colonne che mostrano i criteri specificati (media e deviazione standard) e il valore Lambda.
- Categorie obiettivo riordinate per migliorare la stabilità.

Tabella Predittori categoriali

La tabella mostra il numero di predittori categoriali:

- Le cui categorie sono state riordinate dalla più bassa alla più alta per migliorare la stabilità.
- Le cui categorie sono state unite per aumentare al massimo l'associazione all'obiettivo.
- Le cui categorie sono state unite per gestire le categorie sparse.
- Esclusi a causa di una scarsa associazione all'obiettivo.
- Esclusi perché erano costanti dopo l'unione.

Se non erano presenti predittori categoriali viene visualizzata una nota a piè di pagina.

Tabella Predittori continui

In questo caso le tabelle sono due. La prima visualizza uno dei seguenti numeri di trasformazioni:

- Valori dei predittori trasformati in unità standard. Sono visualizzati inoltre il numero dei predittori trasformati, la media specificata e la deviazione standard.
- Valori dei predittori associati a un intervallo comune. Sono visualizzati inoltre il numero di predittori trasformati mediante una trasformazione Min/Max e i valori minimi e massimi specificati.
- Valori di predittori categorizzati e il numero di predittori categorizzati.

La seconda tabella riporta i dettagli di creazione dello spazio dei predittori, visualizzati sotto forma di numero di predittori:

- Creati.
- Esclusi a causa di una scarsa associazione all'obiettivo.
- Esclusi perché erano costanti dopo la categorizzazione.
- Esclusi perché erano costanti dopo la creazione.

Se l'input non includeva predittori continui viene visualizzata una nota a piè di pagina.

Trasforma all'indietro i punteggi

Se un obiettivo è stato trasformato da ADP, i modelli successivi costruiti utilizzando l'obiettivo trasformato segnano le unità trasformate. Per interpretare e utilizzare questi risultati è necessario riconvertire il valore previsto nella scala originale.

1. Per retrotrasformare i punteggi, dai menu scelgono:

   Trasforma > **Preparare i dati per la modellazione** > **Backtransform Scores ...**

2. Selezionare un campo da backtrasformare. Questo campo deve contenere valori preprevisti del target trasformato.
3. Specificare un suffisso per il nuovo campo. Questo nuovo campo conterrà valori previsti dal modello nella scala originale dell'obiettivo non trasformato.
4. Specificare la posizione del file XML contenente le trasformazioni ADP. Questo dovrebbe essere un file salvato dai dialoghi Interactive o Automatic Data Preparation. Per ulteriori informazioni, consultare la sezione “Applicazione e Saving Trasformazioni ” a pagina 12.

Identifica casi insoliti

La Procedura di rilevamento delle anomalie cerca casi insoliti in base alle deviazioni dalle norme dei loro gruppi di cluster. La procedura è progettata per rilevare rapidamente casi insoliti per scopi di controllo dei dati nel passo di analisi dei dati esplorativi, prima di qualsiasi analisi di dati inferenziale. Questo algoritmo è progettato per il rilevamento di anomalie generiche; cioè la definizione di un caso anomalo non è specifica per nessuna applicazione particolare, come il rilevamento di schemi di pagamento insoliti nell'industria sanitaria o il rilevamento del riciclaggio di denaro nel settore delle finanze, in cui la definizione di un'anomalia può essere ben definita.

Esempio. Un analista di dati assunto per costruire modelli predittivi per gli esiti del trattamento degli ictus si preoccupa della qualità dei dati perché tali modelli possono essere sensibili a osservazioni insolite. Alcune di queste osservazioni stravaganti rappresentano casi davvero unici e sono quindi inadeguate per la previsione, mentre altre osservazioni sono causate da errori di inserimento dati in cui i valori sono tecnicamente "corretti" e quindi non possono essere catturati dalle procedure di convalida dei dati. L'Identificazione dei casi insoliti trova e riporta questi outlier in modo che l'analista possa decidere come gestirli.

Statistiche. La procedura produce peer group, norme di gruppo peer per variabili continue e categoriali, indici di anomalia basati su deviazioni dalle norme del gruppo peer, e valori di impatto variabili per le variabili che maggiormente contribuiscono ad un caso considerato insolita.

Considerazioni sui dati

Dati. Questa procedura può essere utilizzata sia con le variabili continue sia con le variabili categoriali. Ogni riga rappresenta un'osservazione distinta e ogni colonna rappresenta una variabile distinta su cui si basano i gruppi peer. Una variabile di identificazione del caso può essere disponibile nel file dei dati per l'emissione di marcatura, ma non verrà utilizzata nell'analisi. I valori mancanti sono consentiti. La variabile di peso, se specificata, viene ignorata.

Il modello di rilevamento può essere applicato ad un nuovo file di dati di test. Gli elementi dei dati di prova devono essere gli stessi degli elementi dei dati formativi. E, a seconda delle impostazioni dell'algoritmo, la gestione del valore mancante che viene utilizzata per creare il modello può essere applicata al file dei dati di prova prima di scorrere.

Ordine dei casi. Si noti che la soluzione può dipendere dall'ordine dei casi. Per ridurre al minimo gli effetti dell'ordine, disporre i casi in ordine casuale. Per verificare la stabilità di una determinata soluzione, è possibile ottenere diverse soluzioni con casi ordinati in diversi ordini casuali. Nelle situazioni con dimensioni file estremamente grandi, le esecuzioni multiple possono essere eseguite con un campione di casi ordinati in diversi ordini casuali.

Ipotesi. L'algoritmo presuppone che tutte le variabili siano non costanti e indipendenti e che nessun caso abbia valori mancanti per nessuna delle variabili di input. Ogni variabile continua è ipotizzata per avere una distribuzione normale (gaussiana) e ogni variabile categoriale è ipotizzata per avere una distribuzione multinomiale. I test interni empirici indicano che la procedura è abbastanza robusta per le violazioni sia dell'assunzione di indipendenza che delle ipotesi distributive, ma attenzione a quanto siano soddisfatte queste ipotesi.

Per identificare casi insoliti

1. Dai menu, scegliere:

   Dati > **Identificare Casi Insoliti ...**

2. Selezionare almeno una variabile di analisi.
3. Opzionalmente, scegliere una variabile identificativo del caso da utilizzare in output di etichettatura.

Campi con livello di misurazione sconosciuto

L'avviso Livello di misurazione viene visualizzato quando il livello di misurazione di una o più variabili (campi) del dataset è sconosciuto. Poiché influisce sul calcolo dei risultati di questa procedura, il livello di misurazione deve essere definito per tutte le variabili.

Esegui scansione dati. Legge i dati del dataset attivo e assegna un livello di misurazione predefinito a tutti i campi con livello di misurazione sconosciuto. Con dataset di grandi dimensioni, questa operazione può richiedere del tempo.

Assegna manualmente. Apre una finestra di dialogo che elenca tutti i campi con livello di misurazione sconosciuto, mediante la quale è possibile assegnare un livello di misurazione a questi campi. Il livello di misurazione si può assegnare anche nella Vista variabile dell'Editor dei dati.

Dal momento che il livello di misurazione è importante per questa procedura, è possibile accedere alla finestra di dialogo per la sua esecuzione solo quando per tutti i campi è stato definito un livello di misurazione.

Identificazione Casi Insoliti Output

Elenco di casi insoliti e motivi per cui sono considerati insoliti. Questa opzione produce tre tabelle:

- L'elenco degli indici dei casi di anomalia visualizza i casi che vengono identificati come insoliti e visualizza i relativi valori di indice di anomalia.
- Il caso di anomalia caso ID peer list visualizza casi insoliti e informazioni riguardanti i rispettivi gruppi peer peer.
- L'elenco delle ragioni di anomalia visualizza il numero del caso, la variabile motivo, il valore di impatto variabile, il valore della variabile e la norma della variabile per ogni motivo.

Tutte le tabelle sono ordinate per indice di anomalia in ordine decrescente. Inoltre, gli ID dei casi vengono visualizzati se la variabile identificativo del caso è specificata nella scheda Variabili.

Riepiloghi. I controlli di questo gruppo permettono di generare riassunti delle distribuzioni.

- **Norme di gruppo peer.** Questa opzione visualizza la tabella delle norme variabili continue (se si utilizza qualsiasi variabile continua nell'analisi) e la tabella delle norme di variabili categoriali (se nell'analisi viene utilizzata qualsiasi variabile categoriale). La tabella delle norme variabili continue visualizza la deviazione media e standard di ogni variabile continua per ogni gruppo peer. La tabella delle norme variabili categoriali visualizza la modalità (categoria più popolare), la frequenza e la percentuale di frequenza di ogni variabile categoriale per ogni gruppo peer. La media di una variabile continua e la modalità di una variabile categoriale sono utilizzati come valori norm nell'analisi.
- **Indici di anomalia.** Il riepilogo dell'indice di anomalia visualizza le statistiche descrittive per l'indice di anomalia dei casi che sono identificati come i più insoliti.
- **Ricorrenza per variabile di analisi.** Per ogni motivo, la tabella visualizza la frequenza e la percentuale di frequenza di ogni ricorrenza di ogni variabile come motivo. La tabella riporta anche le statistiche descrittive dell'impatto di ciascuna variabile. Se il numero massimo di motivi è impostato a 0 nella scheda Opzioni, questa opzione non è disponibile.
- **Casi elaborati.** Il riepilogo dell'elaborazione dei casi visualizza i conteggi e le percentuali di conteggio per tutti i casi nel dataset attivo, i casi inclusi ed esclusi nell'analisi, e i casi in ogni gruppo peer.

Identificazione Casi Insoliti Salvi

Salva Variabili. I controlli di questo gruppo consentono di salvare le variabili modello nel dataset attivo. Puoi anche scegliere di sostituire le variabili esistenti i cui nomi entrano in conflitto con le variabili da salvare.

- **Indice di anomalia.** Salva il valore dell'indice di anomalia per ogni caso a una variabile con il nome specificato.

- **Gruppi peer.** Salva l'ID del gruppo peer, il conteggio dei casi e la dimensione in percentuale per ogni caso a variabili con il rootname specificato. Ad esempio, se viene specificato il rootname *Peer* , vengono generate le variabili *Peerid*, *PeerSize*e *PeerPctSize* . *Peerid* è l'ID del gruppo peer del caso, *PeerSize* è la dimensione del gruppo e *PeerPctSize* è la dimensione del gruppo come percentuale.
- **Motivi.** Salva serie di variabili di ragionamento con il rootname specificato. Una serie di variabili di ragionamento consiste nel nome della variabile come la ragione, la sua misura d'impatto variabile, il proprio valore e il valore norm. Il numero di serie dipende dal numero di motivi richiesti nella scheda Opzioni. Ad esempio, se viene specificato il rootname *Reason* , vengono generate le variabili *RagionVar_k*, *RagionMeasure_k*, *Ragionvalue_k*e *Ragionnorm_k* , dove *k* è il motivo *k*. Questa opzione non è disponibile se il numero di motivi è impostato su 0.

Export Modello file. Consente di salvare il modello in formato XML.

Identificazione Casi Insoliti Valori mancanti

La scheda Valori mancanti viene utilizzata per controllare la gestione dei valori mancanti di utenza e di sistema.

- **Escludere i valori mancanti da analisi.** I casi con valori mancanti sono esclusi dall'analisi.
- **Include i valori mancanti nell'analisi.** I valori mancanti di variabili continue sono sostituibili con i relativi grand mezzi e le categorie mancanti di variabili categoriali sono raggruppate e trattate come una categoria valida. Le variabili elaborate vengono poi utilizzate nell'analisi. Facoltativamente, è possibile richiedere la creazione di una variabile aggiuntiva che rappresenti la proporzione di variabili mancanti in ogni caso e utilizzare quella variabile nell'analisi.

Identificare Opzioni di casi insoliti

Criteri per l'identificazione dei casi insoliti. Queste selezioni determinano quanti casi sono inclusi nella lista delle anomalie.

- **Percentuale di casi con valori di indice di anomalia più elevati.** Specificare un numero positivo inferiore o uguale a 100.
- **Numero fisso di casi con valori di indice di anomalia più elevati.** Specificare un intero positivo inferiore o uguale al numero totale di casi nel dataset attivo che vengono utilizzati nell'analisi.
- **Identificare solo i casi il cui valore di indice di anomalia soddisfa o supera un valore minimo.** Specificare un numero non negativo. Un caso è considerato anomalo se il suo valore di indice di anomalia è più grande o uguale al punto di demarcazione specificato. Questa opzione viene utilizzata insieme alla **Percentuale di casi** e **Numero di casi fisso** . Ad esempio, se si specifica un numero fisso di 50 casi e un valore di cutoff di 2, l'elenco delle anomalie consisterà, al massimo, 50 casi, ognuno con un valore di indice di anomalia maggiore o uguale a 2.

Numero di gruppi Peer. La procedura ricercherà il maggior numero di gruppi peer tra i valori minimi e massimi specificati. I valori devono essere interi positivi e il minimo non deve superare il massimo. Quando i valori specificati sono uguali, la procedura assume un numero fisso di gruppi peer.

Nota: a seconda della quantità di variazione dei tuoi dati, ci possono essere situazioni in cui il numero di peer group che i dati possono supportare è inferiore al numero specificato come minimo. In una situazione del genere, la procedura può produrre un numero minore di gruppi di pari.

Numero massimo di Motivi. Un motivo consiste nella misura d'impatto variabile, il nome variabile per questo motivo, il valore della variabile e il valore del corrispondente gruppo peer. Specificare un intero non negativo; se questo valore equivale o supera il numero di variabili elaborate che vengono utilizzate nell'analisi, vengono mostrate tutte le variabili.

RILEVTANOMALY Comando Funzioni aggiuntive

Il linguaggio della sintassi dei comandi consente inoltre di:

- Omesso alcune variabili nel dataset attivo dall'analisi senza specificare esplicitamente tutte le variabili di analisi (utilizzando il comando EXCEPT).

- Specificare una regolazione per bilanciare l'influenza delle variabili continue e categoriali (utilizzando la parola chiave `MLWEIGHT` nel comando `CRITERIA`).

Vedere *Command Syntax Reference* per informazioni dettagliate sulla sintassi.

Raccolta ottimale

La Procedura di Binning ottimale discreta una o più variabili di scala (indicate da henceforth come **binning input variabili**) distribuendo i valori di ciascuna variabile in bins. La formazione di Bin è ottimale rispetto a una variabile guida categoriale che "supervisori" il processo di binning. I bidoni possono quindi essere utilizzati invece dei valori dati originali per ulteriori analisi.

Esempi. Riducendo il numero di valori distinti una variabile prende un numero di utilizzi, tra cui:

- Requisiti dati di altre procedure. Le variabili discretizzate possono essere trattate come categoriali per l'utilizzo in procedure che richiedono variabili categoriali. Ad esempio, la procedura Crosstabs richiede che tutte le variabili siano categoriali.
- Privacy dei dati. Segnalare valori binati invece di valori reali può aiutare a salvaguardare la privacy delle vostre fonti di dati. La Procedura di Binning ottimale può orientare la scelta dei bidoni.
- Prestazioni di velocità. Alcune procedure sono più efficienti quando si lavora con un numero ridotto di valori distinti. Ad esempio, la velocità di Regressione Logistica Multinomiale può essere migliorata utilizzando variabili discretizzate.
- Scoprire separazione completa o quasi completa dei dati.

Ottimale contro il Binning visivo. Le finestre di dialogo di Visual Binning offrono diversi metodi automatici per la creazione di bins senza l'utilizzo di una variabile guida. Queste regole "non supervisate" sono utili per produrre statistiche descrittive, come le tabelle di frequenza, ma Ottimale Binning è superiore quando il tuo obiettivo finale è quello di produrre un modello predittivo.

Output. La procedura produce tabelle di cutpoint per i bidoni e le statistiche descrittive per ogni variabile di input di binning. Inoltre, è possibile salvare nuove variabili nel dataset attivo contenente i valori binati delle variabili di input di binning e salvare le regole di binning come sintassi di comando da utilizzare in discretizing nuovi dati.

Considerazioni Sui Dati Di Binning ottimali

Dati. Questa procedura prevede che le variabili di input di binning siano di scala, variabili numeriche. La variabile guida deve essere categoriale e può essere stringa o numerica.

Per ottenere un binning ottimale

1. Dai menu, scegliere:

   Trasformazione > **Binning ottimale ...**
2. Selezionare una o più variabili di input di binning.
3. Selezionare una variabile guida.

Le variabili contenenti i valori dei dati binati non vengono generate per impostazione predefinita. Utilizzare la scheda Salva per salvare queste variabili.

Output di Binning ottimale

La scheda Output controlla la visualizzazione dei risultati.

- **Endpoint per i bidoni.** Visualizza la serie di endpoint per ogni variabile di input di binning.
- **Statistiche descrittive per le variabili che vengono binate.** Per ogni variabile di input di binning, questa opzione visualizza il numero di casi con valori validi, il numero di casi con valori mancanti, il numero di valori validi distinti e i valori minimi e massimi. Per la variabile guida, questa opzione visualizza la distribuzione di classe per ogni variabile di input di binning correlata.
- **Modello di entropia per le variabili che vengono binate.** Per ogni variabile di input di binning, questa opzione visualizza una misura dell'accuratezza predittiva della variabile rispetto alla variabile guida.

Salvataggio ottimale Binning

Salvataggio delle variabili in Dataset attivo. Le variabili contenenti i valori dei dati binati possono essere utilizzate al posto delle variabili originali in ulteriori analisi.

Salva regole di raccolta come Syntax. Genera sintassi di comando che può essere utilizzata per bin altri datasets. Le regole di ricovero si basano sui tagli determinati dall'algoritmo di binning.

Valori di Binning mancanti ottimali

La scheda Valori mancanti specifica se i valori mancanti vengono gestiti utilizzando l'eliminazione listwise o pairwise. I valori mancanti definiti dall'utente vengono sempre considerati come non validi. Quando si recupera i valori variabili originali in una nuova variabile, i valori mancanti dell'utente vengono convertiti in sistema - mancante.

- **Pairwise.** Questa opzione opera su ogni coppia di variabili di input e binning. La procedura farà uso di tutti i casi con valori non mancanti sulla variabile di input guida e binning.
- **Listwise** Questa opzione funziona in tutte le variabili specificate sulla scheda Variabili. Se manca una variabile per un caso, l'intero caso è escluso.

Opzioni di binning ottimali

Preelaborazione. "Pre - Binning" binning input variabili con molti valori distinti può migliorare il tempo di elaborazione senza un grande sacrificio nella qualità dei bidoni finali. Il numero massimo di bidoni dà un limite superiore sul numero di bidoni creati. Così, se si specifica 1000 come il massimo ma una variabile di input di binning ha meno di 1000 valori distinti, il numero di bidoni preelaborati creati per la variabile di input di binning egualerà il numero di valori distinti nella variabile di input di binning.

Bins scarsamente popolati. Occasionalmente la procedura può produrre bidoni con pochissimi casi. La seguente strategia cancella questi pseudo tagli:

Per una determinata variabile, supponiamo che l'algoritmo abbia trovato n_{finali} tagli e quindi $n_{finale}+1$ bins. Per i bidoni $i = 2, ..., n_{finale}$ (il secondo bidone più basso attraverso il secondo vassoio più alto), calcola

$$\frac{sizeof(b_i)}{\min(sizeof(b_{i-1}), sizeof(b_{i+1}))}$$

dove *sizeof (b)* è il numero di casi nel bidone.

Quando questo valore è inferiore alla soglia di unione specificata, b_i è considerato scarsamente popolato e viene unito a b_{i-1} o b_{i+1}, qualunque sia l'entropia di informazioni di classe inferiore.

La procedura fa passare un singolo passaggio attraverso i bidoni.

Bin Endpoints. Questa opzione specifica come viene definito il limite inferiore di un intervallo. Dato che la procedura determina automaticamente i valori dei tagli, si tratta in larga parte di una questione di preferenza.

Primo (Minimo) / Ultimo (Massimo) Bin. Queste opzioni specificano come sono definiti i cutpoint minimi e massimi per ogni variabile di input di binning. Generalmente la procedura presuppone che le variabili di input di binning possano assumere qualsiasi valore sulla linea del numero reale, ma se si ha qualche motivo teorico o pratico per limitare la gamma, è possibile vincolarlo dai valori più bassi / massimi.

BINNING OTTIMALE Comando Funzioni Aggiuntive

Il linguaggio della sintassi dei comandi consente inoltre di:

- Eseguire il binning non supervisore tramite il metodo delle frequenze uguali (utilizzando il comando `CRITERIA`).

Vedere *Command Syntax Reference* per informazioni dettagliate sulla sintassi.

Informazioni particolari

Queste informazioni sono state sviluppate per prodotti e servizi offerti negli Stati Uniti. Questo materiale potrebbe essere disponibile da IBM in altre lingue. Tuttavia, all'utente potrebbe essere richiesto di possedere una copia del prodotto o una versione del prodotto in tale lingua per accedervi.

IBM può non offrire i prodotti, i servizi o le funzioni presentati in questo documento in altri paesi. Consultare il proprio rappresentate locale IBM per informazioni sui prodotti ed i servizi attualmente disponibili nella propria zona. Qualsiasi riferimento ad un prodotto, programma o servizio IBM non implica o intende dichiarare che solo quel prodotto, programma o servizio IBM può essere utilizzato. In sostituzione a quelli forniti da IBM, è possibile usare prodotti, programmi o servizi funzionalmente equivalenti che non comportino violazione dei diritti di proprietà intellettuale o di altri diritti di IBM. Tuttavia, è responsabilità dell'utente valutare e verificare il funzionamento di qualsiasi prodotto, programma o servizio non IBM.

IBM può avere applicazioni di brevetti o brevetti in corso relativi all'argomento descritto in questo documento. La fornitura di questa documentazione non concede alcuna licenza su questi brevetti. È possibile inviare per iscritto richieste di licenze a:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
USA

Per richieste di licenze relative ad informazioni double-byte (DBCS), contattare il Dipartimento di Proprietà Intellettuale IBM nel proprio paese o inviare richieste per iscritto a:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan

IBM (INTERNATIONAL BUSINESS MACHINES CORPORATION) FORNISCE LA PRESENTE PUBBLICAZIONE "NELLO STATO IN CUI SI TROVA" SENZA GARANZIE DI ALCUN TIPO, ESPRESSE O IMPLICITE, IVI INCLUSE, A TITOLO DI ESEMPIO, GARANZIE IMPLICITE DI NON VIOLAZIONE, DI COMMERCIABILITÀ E DI IDONEITÀ PER UNO SCOPO PARTICOLARE. Alcune giurisdizioni non consentono la rinuncia ad alcune garanzie espresse o implicite in determinate transazioni, pertanto, la presente dichiarazione può non essere applicabile.

Questa pubblicazione potrebbe contenere imprecisioni tecniche o errori tipografici. Le modifiche vengono apportate periodicamente alle informazioni qui contenute; questi cambiamenti saranno incorporati nelle nuove edizioni della pubblicazione. IBM si riserva il diritto di apportare miglioramenti e/o modifiche al prodotto o al programma descritto nel manuale in qualsiasi momento e senza preavviso.

I riferimenti in queste informazioni a siti Web non IBM vengono forniti solo per comodità e non implicano in alcun modo l'approvazione di tali siti web. I matteriali disponibili su tali siti Web non fanno parte del materiale relativo a questo prodotto IBM e l'utilizzo di questi è a discrezione dell'utente.

IBM può utilizzare o distribuire qualsiasi informazione fornita in qualsiasi modo ritenga appropriato senza incorrere in alcun obbligo verso l'utente.

Coloro che detengano la licenza su questo programma e desiderano avere informazioni su di esso allo scopo di consentire: (i) uno scambio di informazioni tra programmi indipendenti ed altri (compreso questo) e (ii) l'utilizzo reciproco di tali informazioni, dovrebbe rivolgersi a:

IBM Director of Licensing
IBM Corporation

North Castle Drive, MD-NC119
Armonk, NY 10504-1785
USA

Tali informazioni potrebbero essere disponibili secondo termini e condizioni appropriati compreso, in alcuni casi, il pagamento di un corrispettivo.

Il programma concesso in licenza descritto nel presente documento e tutto il materiale concesso in licenza disponibile sono forniti da IBM in base alle clausole dell'Accordo per Clienti IBM (IBM Customer Agreement), dell'IBM IPLA (IBM International Program License Agreement) o qualsiasi altro accordo equivalente tra le parti.

I dati delle prestazioni e gli esempi client citati vengono presentati solo a scopo illustrativo. Gli effettivi risultati delle prestazioni possono variare in base alle configurazioni e alle condizioni operative specifiche.

Le informazioni relative a prodotti non IBM sono ottenute dai fornitori di quei prodotti, dagli annunci pubblicati i da altre fonti disponibili al pubblico. IBM non ha testato quei prodotti e non può confermarne la precisione della prestazione, la compatibilità o qualsiasi altro reclamo relativo ai prodotti non IBM. Le domande sulle funzionalità dei prodotti non IBM devono essere 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.

Queste informazioni contengono esempi di dati e report utilizzati nelle operazioni aziendali quotidiane. Pertanto, per maggiore completezza, gli esempi includono nomi di persone, società, marchi e prodotti. Tutti i nomi contenuti nel manuale sono fittizi e ogni riferimento a persone o aziende reali è puramente casuale.

LICENZA DI COPYRIGHT:

Queste informazioni contengono programmi campione di applicazione nella lingua di origine, i quali illustrano le tecniche di programmazione su varie piattaforme operative. È possibile copiare, modificare e distribuire questi programmi di esempio sotto qualsiasi forma senza alcun pagamento a IBM, allo scopo di sviluppare, utilizzare, commercializzare o distribuire i programmi applicativi in conformità alle API (application programming interface) a seconda della piattaforma operativa per cui i programmi di esempio sono stati scritti. Questi esempi non sono stati testati approfonditamente tenendo conto di tutte le condizioni possibili. IBM, quindi, non può garantire o sottintendere l'affidabilità, l'utilità o il funzionamento di questi programmi. I programmi di esempio sono forniti "COSÌ COME SONO", senza garanzie di alcun tipo. IBM non intende essere responsabile per alcun danno derivante dall'uso dei programmi di esempio.

Ogni copia o qualsiasi parte di questi programmi di esempio o qualsiasi lavoro derivato, devono contenere le seguenti informazioni relative alle leggi sul diritto d'autore:

Marchi

IBM, il logo IBM e ibm.com sono marchi o marchi registrati di International Business Machines Corp., registrati in molte giurisdizioni in tutto il mondo. Altri nomi di prodotti e servizi possono essere marchi di IBM o di altre società. Un elenco corrente dei marchi IBM è disponibile sul web in "Copyright and trademark information" all'indirizzo www.ibm.com/legal/copytrade.shtml.

Adobe, il logo Adobe, PostScript e il logo PostScript sono marchi o marchi registrati di Adobe Systems Incorporated negli Stati Uniti e/o in altri paesi.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium e Pentium sono marchi o marchi registrati di Intel Corporation o relative controllate negli Stati Uniti e altri paesi.

Linux è un marchio registrato di Linus Torvalds negli Stati Uniti e/o in altri paesi.

Microsoft, Windows, Windows NT e il logo Windows sono marchi di Microsoft Corporation negli Stati Uniti e/o in altri paesi.

UNIX è un marchio della The Open Group negli Stati Uniti e/o negli altri paesi.

Java e tutti i marchi e i logo basati su Java sono marchi o marchi registrati di Oracle e/o associate.

Indice analitico

S

T

V