IBM SPSS Data Preparation 19

Note: Before using this information and the product it supports, read the general information under Notices a pag. 148.

Prefazione

IBM® SPSS® Statistics è un sistema completo per l'analisi dei dati. Il modulo aggiuntivo opzionale Data Preparation include le tecniche di analisi aggiuntive descritte nel presente manuale. Il modulo aggiuntivo Data Preparation deve essere usato con il modulo Core SPSS Statistics in cui è completamente integrato.

Informazioni su SPSS Inc., una società del gruppo IBM

SPSS Inc., una società del gruppo IBM, è fornitore leader mondiale nel settore del software e delle soluzioni per l'analisi predittiva. L'offerta completa dei prodotti dell'azienda (raccolta di dati, statistica, modellazione e distribuzione) consente di acquisire i comportamenti e le opinioni delle persone, prevedere i risultati delle future interazioni con i clienti ed elaborare questi dati integrando le analitiche nelle procedure aziendali. Le soluzioni SPSS Inc. consentono la gestione di attività interconnesse all'interno dell'intera organizzazione, con particolare attenzione alla convergenza di analitiche, architettura IT e procedure aziendali. Clienti commerciali, istituzionali e accademici di tutto il mondo si affidano alla tecnologia SPSS Inc. ottenendo un vantaggio competitivo in termini di attrazione, mantenimento e ampliamento della base clienti, riducendo al contempo frodi e rischi. SPSS Inc. è stata acquisita da IBM nell'ottobre 2009. Per ulteriori informazioni, visitare il sito *http://www.spss.com*.

Supporto tecnico

Ai clienti che richiedono la manutenzione, viene messo a disposizione un servizio di supporto tecnico. I clienti possono contattare il supporto tecnico per richiedere assistenza per l'utilizzo dei prodotti SPSS Inc. o per l'installazione di uno degli ambienti hardware supportati. Per il supporto tecnico, visitare il sito Web di SPSS Inc. all'indirizzo *http://support.spss.com* o contattare la filiale del proprio paese indicata nel sito Web all'indirizzo *http://support.spss.com/default.asp?refpage=contactus.asp*. Ricordare che durante la richiesta di assistenza sarà necessario fornire i dati di identificazione personali, i dati relativi alla propria società e il numero del contratto di manutenzione.

Servizio clienti

Per informazioni sulla spedizione o sul proprio account, contattare la filiale nel proprio paese, indicata nel sito Web all'indirizzo *http://www.spss.com/worldwide*. Tenere presente che sarà necessario fornire il numero di serie.

Corsi di formazione

SPSS Inc. organizza corsi di formazione pubblici e onsite che includono esercitazioni pratiche. Tali corsi si terranno periodicamente nelle principali città. Per ulteriori informazioni sui corsi, contattare la filiale nel proprio paese, indicata nel sito Web all'indirizzo *http://www.spss.com/worldwide*.

Pubblicazioni aggiuntive

I documenti *SPSS Statistics: Guide to Data Analysis*, *SPSS Statistics: Statistical Procedures Companion* e *SPSS Statistics: Advanced Statistical Procedures Companion*, scritti da Marija Norušis e pubblicati da Prentice Hall sono disponibili come materiale supplementare consigliato. Queste pubblicazioni descrivono le procedure statistiche nei moduli SPSS Statistics Base, Advanced Statistics e Regression. Utili sia come guida iniziale all'analisi dei dati che per applicazioni avanzate, questi manuali consentono di ottimizzare l'utilizzo delle funzionalità presenti nell'offerta IBM® SPSS® Statistics. Per ulteriori informazioni, inclusi contenuti delle pubblicazioni e capitoli di esempio, visitare il sito Web dell'autrice: *http://www.norusis.com*

Contenuto

Parte I:
Manuale dell'utente

Capitolo
1

Introduzione al modulo Data Preparation

La disponibilità di sistemi più potenti permette di gestire un maggiore volume di dati, il che significa che vengono raccolti sempre più dati. — Tuttavia, l'utilizzo di un maggior numero di casi e variabili incrementa anche il numero di errori di inserimento. Poiché questi errori influiscono significativamente sulle previsioni dei modelli predittivi, che rappresenta l'obiettivo finale del data warehousing, è indispensabile tenere i dati "puliti". Tuttavia, la quantità di dati archiviati è tale che non è più possibile verificare i casi manualmente. Di qui la necessità di adottare processi automatizzati per la convalida dei dati.

Il modulo aggiuntivo Data Preparation permette di identificare i casi insoliti e i casi, le variabili e i valori dei dati non validi in un insieme di dati attivo e di preparare i dati per la modellazione.

Uso delle procedure di Data Preparation

L'uso delle procedure di $$$Preparazione dati varia in base alle specifiche esigenze. Le operazioni tipiche che vengono effettuate dopo il caricamento dei dati sono le seguenti:

- **Preparazione dei metadati.** Rivedere le variabili e stabilire i valori validi, le etichette e i livelli di misurazione. Quindi, identificare le combinazioni di valori di variabili che non sono possibili e probabilmente codificati in modo errato. Definire le regole di convalida in base a queste informazioni. Sebbene queste operazioni possano richiedere molto tempo, sono molto utili soprattutto se si ha l'esigenza di convalidare regolarmente file di dati con attributi simili.
- **convalida dati.** Eseguire i controlli di base regolari utilizzando regole di convalida definite per identificare i casi, le variabili e i valori dei dati non validi. Identificare e correggere la causa degli eventuali valori non validi. Ciò può richiedere un ulteriore passaggio.
- **Preparazione del modello.** Utilizzare la preparazione automatica dati per ottenere trasformazioni dei campi originali in grado di migliorare la creazione del modello. Identificare i possibili valori statistici anomali che possono provocare problemi in molti modelli predittivi. Alcuni di questi valori dipendono da valori di variabili non valide che non sono stati identificati. Ciò può richiedere un ulteriore passaggio.

Dopo aver "pulito" i dati, è possibile iniziare a creare i modelli utilizzando gli altri moduli aggiuntivi.

Capitolo
2

Regole di convalida

Le regole permettono di stabilire se un caso è valido o meno. Esistono due tipi di regole di convalida:

- **Regole per variabili singole.** Le regole per variabili singole consistono in un insieme fisso di controlli che si applicano a una singola variabile, come i controlli per le variabili fuori intervallo. I valori validi delle regole per variabili singole possono essere espressi come intervallo di valori o come elenco di valori accettabili.
- **Regole per più variabili.** Le regole per più variabili sono regole definite dall'utente che possono essere applicate a singole variabili o a combinazioni di variabili. Le regole per più variabili vengono definite mediante un'espressione logica che contrassegna i valori non validi.

Le regole di convalida vengono salvate nel dizionario dei dati del file di dati. Pertanto, è sufficiente specificare la regola una sola volta per poterla riutilizzare in seguito.

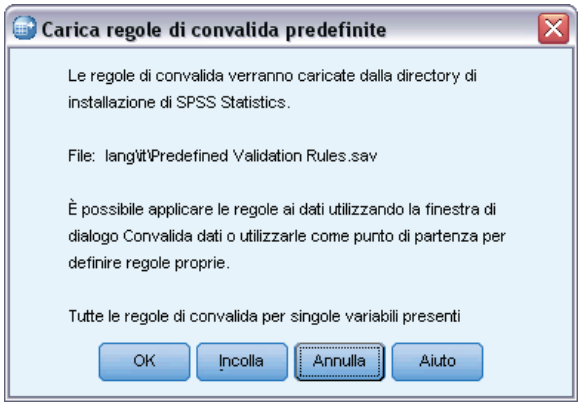
Carica regole di convalida predefinite

Per disporre di un insieme di regole di convalida pronte per l'uso, è sufficiente caricare le regole predefinite da un file di dati esterno incluso durante l'installazione.

Per caricare le regole di convalida predefinite

▶ Dai menu, scegliere:
Dati > Convalida > Carica regole predefinite...

Figura 2-1
Carica regole di convalida predefinite



Ricordare che questo processo elimina tutte le regole per variabili singole del file di dati attivo.

Per caricare le regole da un file di dati, è possibile anche usare Copia proprietà dei dati guidata.

Definisci regole di convalida

La finestra di dialogo Definisci regole di convalida permette di creare e visualizzare regole di convalida per le singole variabili e per più variabili.

Per creare e visualizzare regole di convalida

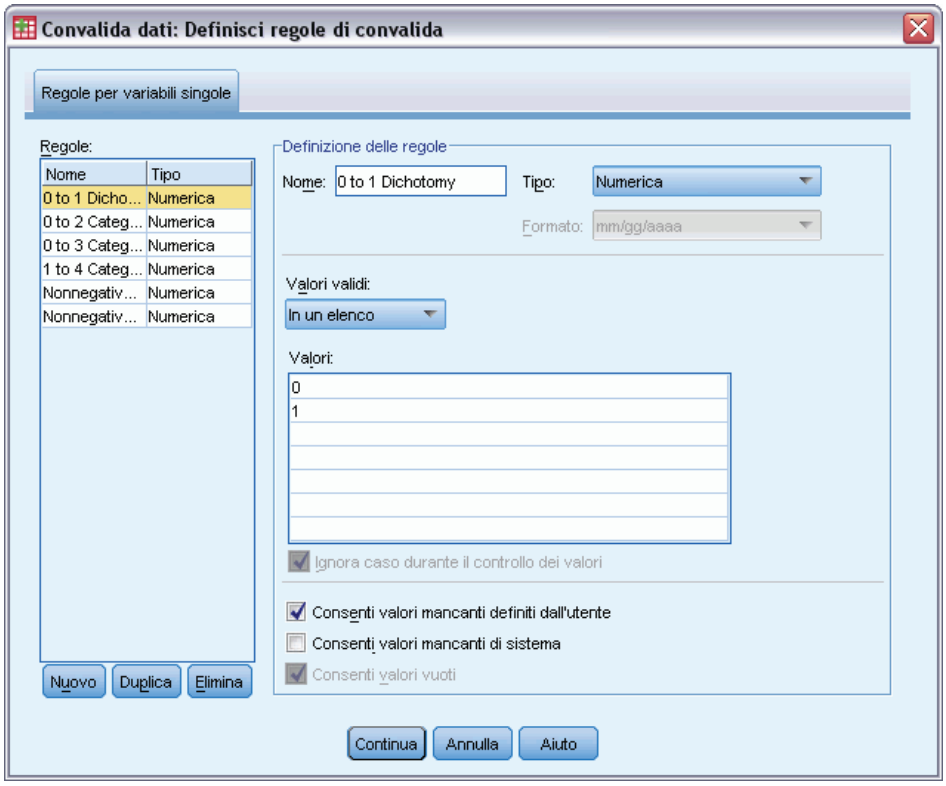
- Dai menu, scegliere:
  Dati > Convalida > Definisci regole...

  Questa finestra di dialogo visualizza automaticamente le regole di convalida per le singole variabili e per più variabili lette dal dizionario dei dati. Se non ci sono regole, viene creata automaticamente una nuova regola per il segnaposto che è possibile modificare in base alle proprie esigenze.

- Selezionare le regole nelle schede Regole per variabili singole e Regole per più variabili per visualizzarne e modificarne le proprietà.

Definisci regole per variabili singole

Figura 2-2
Scheda Regole per variabili singole della finestra di dialogo Definisci regole di convalida



La scheda Regole per variabili singole consente di creare, visualizzare e modificare le regole di convalida per le variabili singole.

Regole. L'elenco visualizza le regole di convalida per variabili singole in base al nome e al tipo di variabile a cui può essere applicata la regola. Se la finestra di dialogo è aperta, visualizza le regole definite nel dizionario dei dati oppure, se non ci sono regole definite, una regola segnaposto chiamata "Regola per variabili singole 1". Sotto all'elenco Regole vengono visualizzati i seguenti pulsanti:

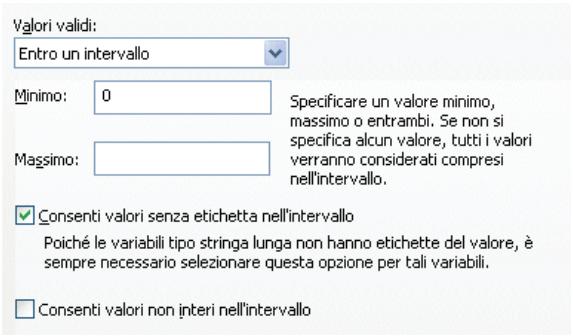
- **Nuovo.** Aggiunge una nuova voce in fondo all'elenco Regole. La regola viene selezionata e viene assegnato il nome "RegVarSingola *n*," dove *n* è un numero intero. Ciò fa sì che il nuovo nome della regola sia univoco nell'ambito delle regole per variabili singole e per più variabili.
- **Duplicata.** Aggiunge una copia della regola selezionata in fondo all'elenco Regole. Il nome della regola viene corretto in modo che risulti univoco nell'ambito delle regole per variabili singole e per più variabili. Ad esempio se si duplica "RegVarSingola 1," il nome della prima regola duplicata sarà "Copia di RegolaVarSingole 1", il secondo "Copia (2) di RegolaVarSingole 1" e così via.
- **Elimina.** Elimina la regola selezionata.

Definizione regole. Questi controlli permettono di visualizzare e impostare le proprietà per ciascuna regola selezionata.

- **Nome.** Il nome della regola deve essere univoco nell'ambito delle regole per variabili singole e più variabili.
- **Tipo.** Indica il tipo di variabile al quale è possibile applicare la regola. Selezionare Numerica, Stringa e Data.
- **Formato.** Permette di selezionare il formato data per le regole da applicare alle variabili data.
- **Valori validi.** Permette di specificare i valori validi come intervallo o elenco di valori.

Questi controlli permettono di specificare un intervallo valido. I valori esterni all'intervallo vengono contrassegnati come non validi.

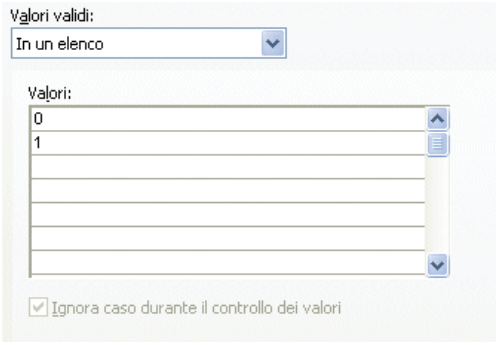
Figura 2-3
Regole per variabili singole: Definizione intervallo



Per specificare un intervallo, immettere un valore minimo o massimo oppure entrambi. I controlli della casella di controllo permettono di contrassegnare i valori non etichettati o non interi compresi nell'intervallo.

Questi controlli permettono di definire un elenco di valori validi. I valori non inclusi nell'elenco vengono contrassegnati come non validi.

Figura 2-4
Regole per variabili singole: Definizione elenco

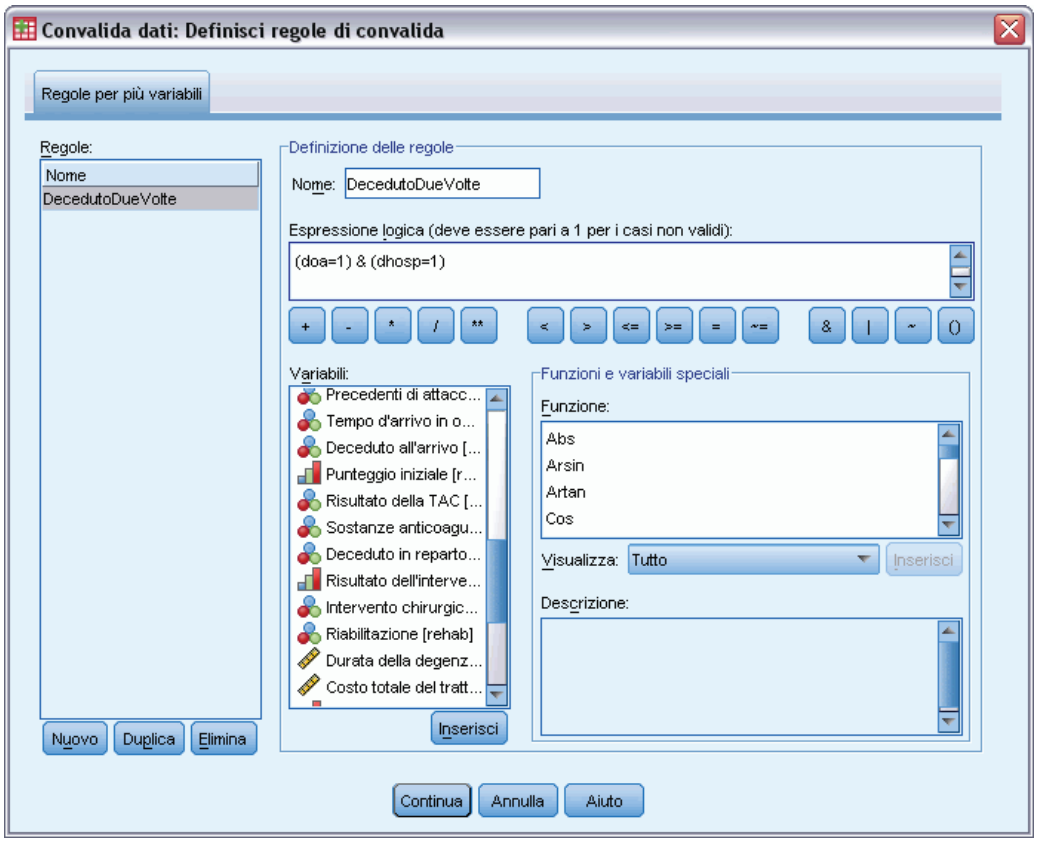


Immettere i valori dell'elenco nella griglia. La casella di controllo determina se i casi sono applicabili o meno quando i valori dei dati stringa vengono confrontati con l'elenco dei valori accettabili.

- **Consenti valori mancanti definiti dall'utente.** Permette di verificare se i valori mancanti definiti dall'utente sono contrassegnati come non validi.
- **Consenti valori mancanti di sistema.** Permette di verificare se i valori mancanti di sistema sono contrassegnati come non validi. Questa opzione non si applica ai tipi di regole per le stringhe.
- **Consenti valori vuoti.** Permette di controllare se i valori stringa vuoti (completamente vuoti) vengono contrassegnati come non validi. Questa opzione non si applica ai tipi di regole diversi dalle stringhe

Definisci regole per più variabili

Figura 2-5
Finestra di dialogo Definisci regole di convalida, scheda Regole per più variabili



La scheda Regole per più variabili consente di creare, visualizzare e modificare le regole di convalida per più variabili.

Regole. Questo elenco visualizza le regole per più variabili per nome. Se la finestra di dialogo è aperta, visualizza la regola segnaposto chiamata "RegPiùVar 1". Sotto all'elenco Regole vengono visualizzati i seguenti pulsanti:

- **Nuovo.** Aggiunge una nuova voce in fondo all'elenco Regole. La regola viene selezionata e viene assegnato il nome "RegPiùVar *n*," dove *n* è un numero intero. Ciò fa sì che il nuovo nome della regola sia univoco nell'ambito delle regole per variabili singole e per più variabili.
- **Duplicata.** Aggiunge una copia della regola selezionata in fondo all'elenco Regole. Il nome della regola viene corretto in modo che risulti univoco nell'ambito delle regole per variabili singole e per più variabili. Ad esempio se si duplica "RegPiùVar 1," il nome della prima regola duplicata sarà "Copia di RegPiùVar 1", il secondo "Copia (2) di RegPiùVar 1" e così via.
- **Elimina.** Elimina la regola selezionata.

Definizione regole. Questi controlli permettono di visualizzare e impostare le proprietà per ciascuna regola selezionata.

- **Nome.** Il nome della regola deve essere univoco nell'ambito delle regole per variabili singole e più variabili.
- **Espressioni logiche.** Rappresenta di fatto la definizione delle regole. È consigliabile codificare l'espressione in modo che i casi non validi vengano valutati come 1.

Creazione di espressioni

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

- È possibile incollare le funzioni o le variabili di sistema normalmente utilizzate selezionando un gruppo nell'elenco delle funzioni e facendo doppio clic su una funzione o variabile nell'elenco delle funzioni e variabili speciali. È possibile anche selezionare la funzione o la variabile, quindi fare clic su Inserisci. Immettere i valori per tutti i parametri contrassegnati con punti interrogativi (solo per le funzioni). Il gruppo di funzioni denominato Tutti contiene un elenco di tutte le funzioni e variabili di sistema disponibili. 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 decimali, usare il punto (.) come separatore decimale.

Capitolo
3

Convalida dati

La finestra di dialogo Convalida dati consente di identificare casi, variabili e valori di dati dubbi e non validi all'interno del file di dati attivo.

Esempio. Un analista di dati deve consegnare al cliente un report mensile sulla soddisfazione dei clienti. I dati ricevuti mensilmente devono essere sottoposti a un controllo qualità per verificare eventuali ID cliente incompleti, valori di variabili fuori intervallo e combinazioni di valori di variabili che vengono comunemente inserite per errore. Dalla finestra di dialogo Convalida dati, l'analista può specificare le variabili che identificano unicamente i clienti, definire le regole per le variabili singole relative agli intervalli di variabili validi e definire le regole per più variabili al fine di individuare le combinazioni impossibili. La procedura crea un report dei casi e delle variabili che presentano dei problemi. Inoltre, i dati presentano gli stessi elementi ogni mese, quindi l'analista può applicare le regole al nuovo file di dati del mese successivo.

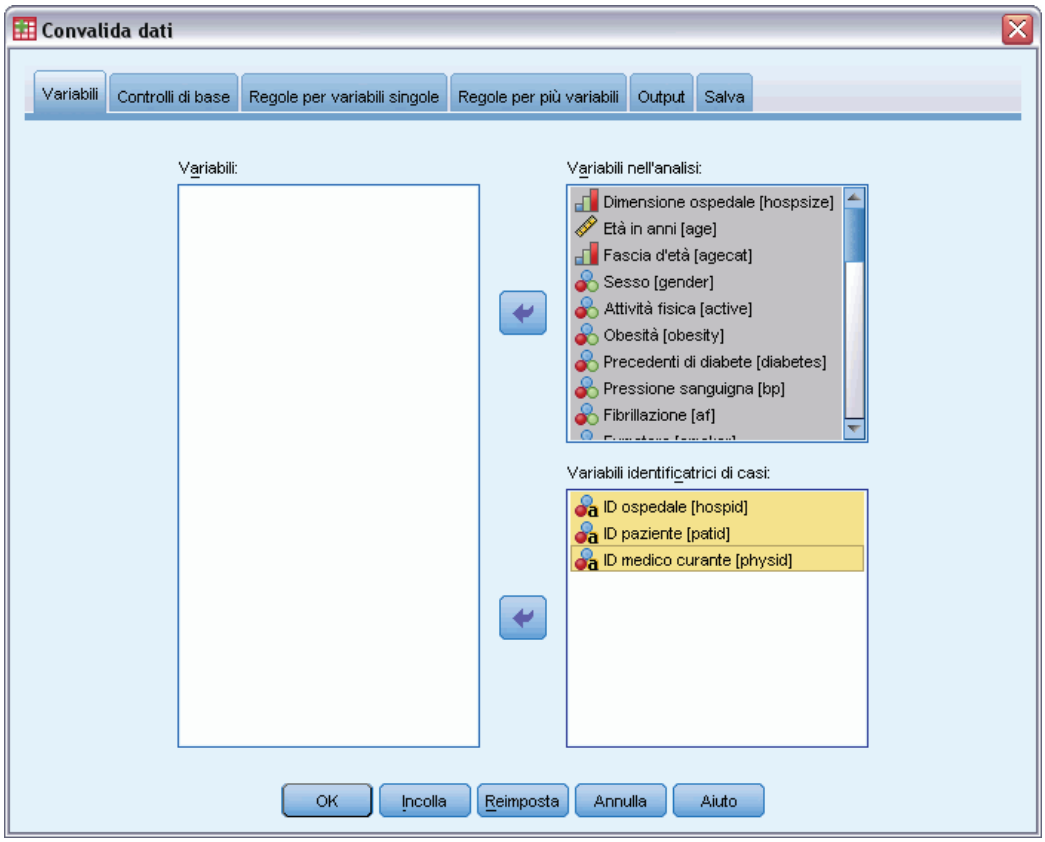
Statistiche. La procedura genera elenchi di variabili, casi e valori di dati che non superano vari controlli, calcola le violazioni delle regole per le variabili singole e per più variabili e produce dei semplici riepiloghi descrittivi delle variabili dell'analisi.

Pesi. La procedura ignora la specifica della variabile di ponderazione e la considera come qualsiasi altra variabile di analisi.

Per eseguire la Convalida dati

- Dai menu, scegliere:
  Dati > Convalida > Convalida dati...

Figura 3-1
Scheda Variabili della finestra di dialogo Convalida dati



▶ Selezionare una o più variabili dell'analisi da convalidare in base ai controlli delle variabili di base o per le regole di convalida per variabili singole.

In alternativa è possibile:

▶ Fare clic sulla scheda Regole per più variabili e applicare una o più regole per più variabili.

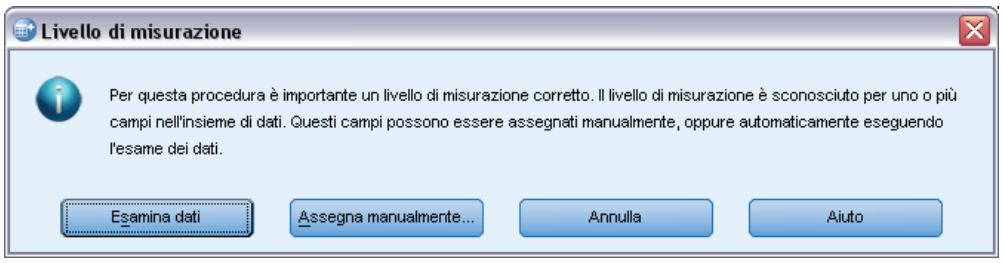
Se lo si desidera, è possibile:

- Selezionare una o più variabili di identificazione dei casi per verificare l'eventuale presenza di ID duplicati o incompleti. È inoltre possibile utilizzare delle variabili ID dei casi per etichettare l'output caso per caso. Se si specificano due o più variabili ID, la combinazione dei rispettivi valori viene considerata come un identificatore di casi.

Campi con livello di misurazione sconosciuto

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

Figura 3-2
Avviso Livello di misurazione

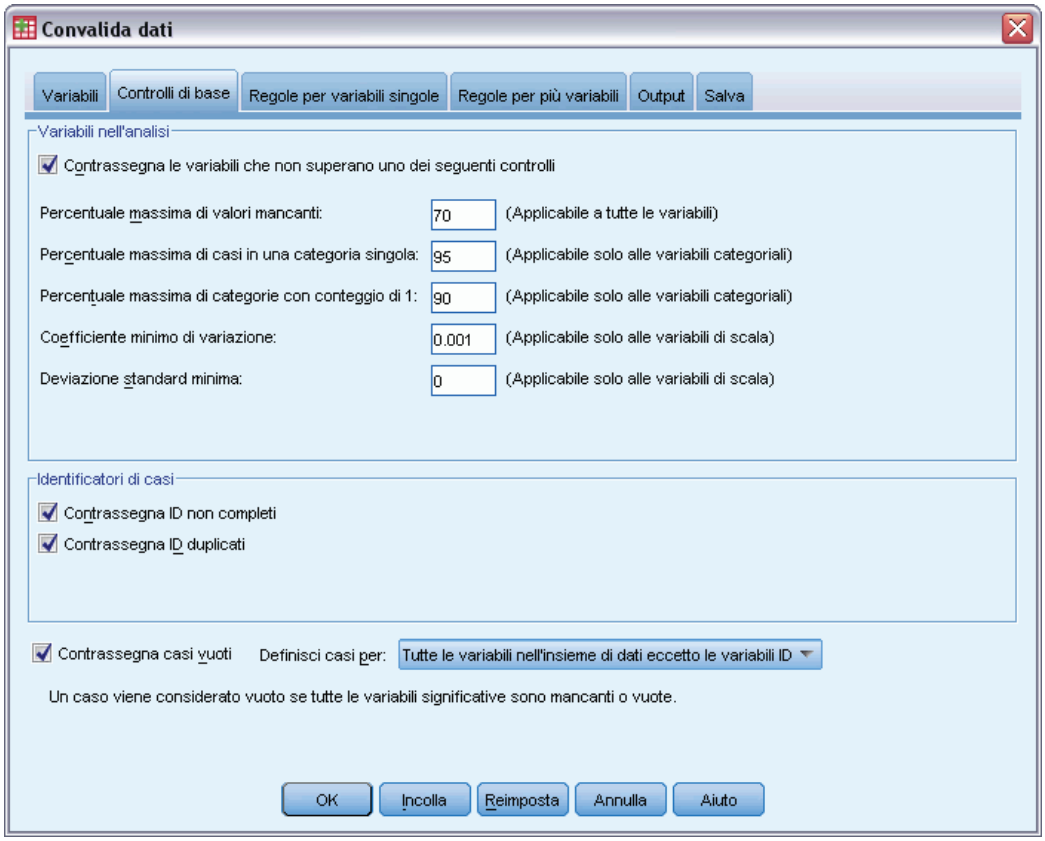


- **Esamina dati.** Legge i dati dell'insieme di dati attivo e assegna un livello di misurazione predefinito a tutti i campi con livello di misurazione sconosciuto. Con insiemi di dati di grandi dimensioni, questa operazione può richiedere del tempo.
- **Assegna manualmente.** Apre una finestra di dialogo 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 Visualizzazione variabili 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.

Controlli di base di Convalida dati

Figura 3-3
Scheda Controlli di base della finestra di dialogo Convalida dati



La scheda Controlli di base consente di selezionare dei controlli di base per le variabili dell'analisi, gli identificatori di casi e per i casi interi.

Variabili dell'analisi. Se nella scheda Variabili si è selezionata una variabile dell'analisi qualsiasi, è possibile applicare i seguenti controlli per verificarne la validità. Dalla rispettiva casella di controllo si attivano e si disattivano i controlli.

- **Percentuale massima di valori mancanti.** Riporta le variabili dell'analisi che presentano una percentuale di valori mancanti superiore al valore specificato. Il valore specificato deve essere un numero positivo inferiore o uguale a 100.
- **Percentuale massima di casi in una categoria singola.** Se qualsiasi variabile dell'analisi è categoriale, l'opzione riporta le variabili dell'analisi categoriali con una percentuale di casi che rappresenta una categoria singola di valori non mancanti superiori al valore specificato. Il valore specificato deve essere un numero positivo inferiore o uguale a 100. La percentuale si basa sui casi con valori non mancanti della variabile.
- **Percentuale massima di categorie con conteggio di 1.** Se qualsiasi variabile dell'analisi è categoriale, l'opzione riporta le variabili dell'analisi categoriali in cui la percentuale di categorie di variabili'i che contengono un solo caso è maggiore del valore specificato. Il valore specificato deve essere un numero positivo inferiore o uguale a 100.

- **Coefficiente minimo di variazione.** Se qualsiasi variabile dell'analisi è di scala, l'opzione riporta le variabili di scala dell'analisi in cui il valore assoluto del coefficiente di variazione è inferiore al valore specificato. L'opzione viene applicata solo alle variabili in cui la media è diversa da zero. Il valore specificato deve essere un numero non negativo. Specificando 0 si disattiva il controllo del coefficiente di variazione.
- **Deviazione standard minima.** Se qualsiasi variabile dell'analisi è di scala, l'opzione riporta le variabili di scala dell'analisi la cui deviazione standard è inferiore al valore specificato. Il valore specificato deve essere un numero non negativo. Specificando 0 si disattiva il controllo della deviazione standard.

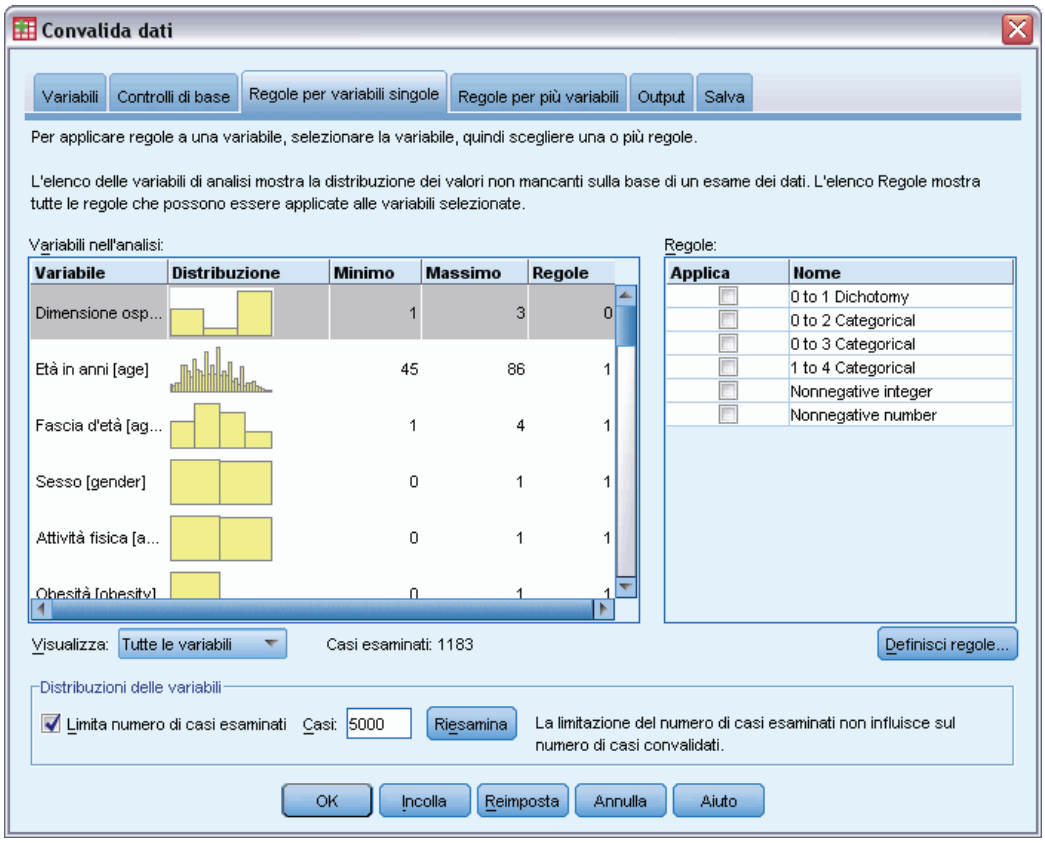
Identificatori di casi. Se nella scheda Variabili si è selezionata una variabile identificatrice di casi qualsiasi, è possibile applicare i seguenti controlli per verificarne la validità.

- **Contrassegna ID non completi.** L'opzione riporta i casi in cui gli identificatori sono incompleti. Per un determinato caso, l'identificatore viene considerato incompleto se il valore di qualsiasi variabile ID è vuoto o mancante.
- **Contrassegna ID duplicati.** L'opzione riporta i casi in cui gli identificatori sono duplicati. Gli identificatori incompleti vengono esclusi dall'insieme di possibili duplicati.

Contrassegna casi vuoti. L'opzione riporta i casi in cui tutte le variabili sono vuote o mancanti. Al fine di identificare i casi vuoti, è possibile utilizzare tutte le variabili nel file (tranne le variabili ID) o le variabili dell'analisi definite nella scheda Variabili.

Regole per variabili singole di Convalida dati

Figura 3-4
Scheda Regole per variabili singole della finestra di dialogo Convalida dati



La scheda Regole per variabili singole visualizza le regole di convalida disponibili per le variabili singole e consente di applicarle alle variabili dell'analisi. Per definire ulteriori regole per variabili singole, fare clic su Definisci regole. Per ulteriori informazioni, vedere l'argomento Definisci regole per variabili singole in il capitolo 2 a pag. 3.

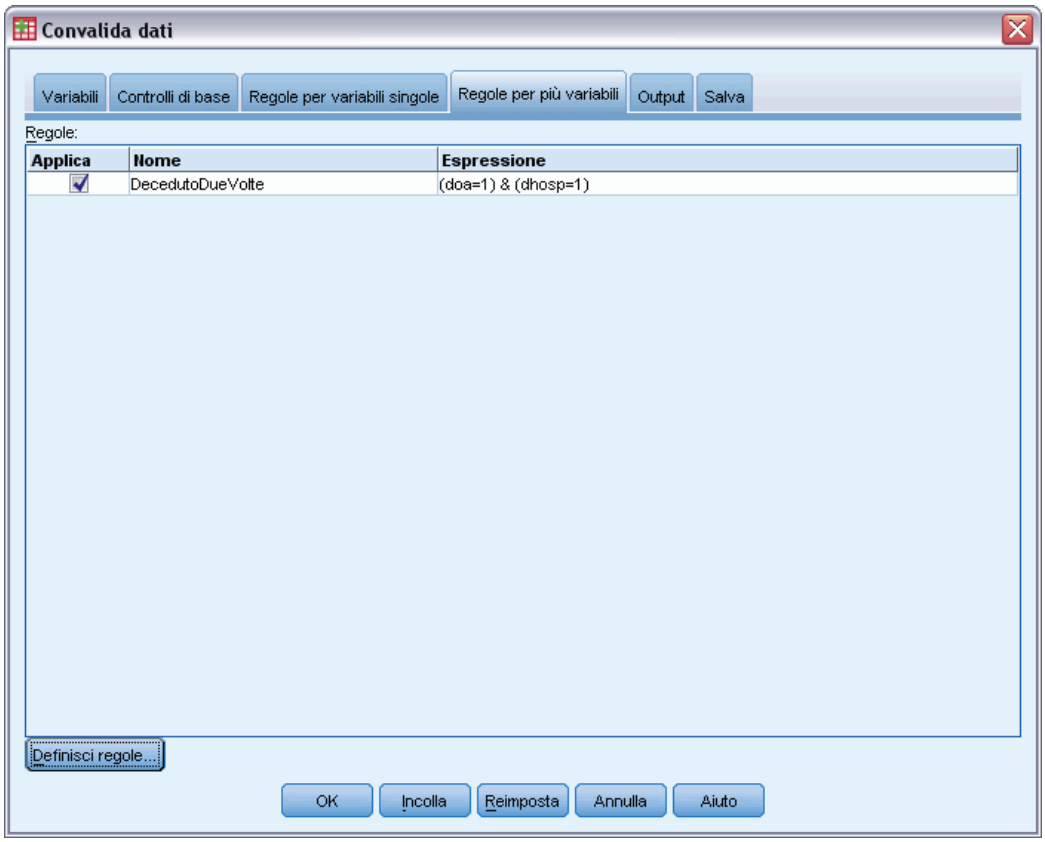
Variabili dell'analisi. L'elenco riporta le variabili dell'analisi, ne riepiloga la distribuzione e mostra il numero di regole applicate a ogni variabile. Si noti che i valori mancanti definiti dall'utente e dal sistema non sono inclusi nei riepiloghi. L'elenco a discesa Visualizza controlla la visualizzazione delle variabili, disponibile per: Tutte le variabili, Variabili numeriche, Variabili stringa e Variabili data.

Regole. Per applicare le regole alle variabili dell'analisi, selezionare una o più variabili e contrassegnare tutte le regole da applicare nell'elenco Regole. L'elenco Regole mostra solo le regole che risultano appropriate per le variabili dell'analisi selezionate. Se per esempio si selezionano le variabili numeriche dell'analisi, vengono riportate solo le regole numeriche; se invece si seleziona una variabile stringa, vengono mostrate solo le regole corrispondenti. Se non vi sono variabili dell'analisi selezionate o se esse presentano dei tipi di dati misti, non viene mostrata alcuna regola.

Distribuzioni delle variabili. I riepiloghi delle distribuzioni riportati nell'elenco Variabili dell'analisi possono basarsi su tutti i casi o su un'analisi dei primi *n* casi, come specificato nella casella di testo Casi. Per aggiornare i riepiloghi delle distribuzioni, fare clic su Riesamina.

Regole per più variabili di Convalida dati

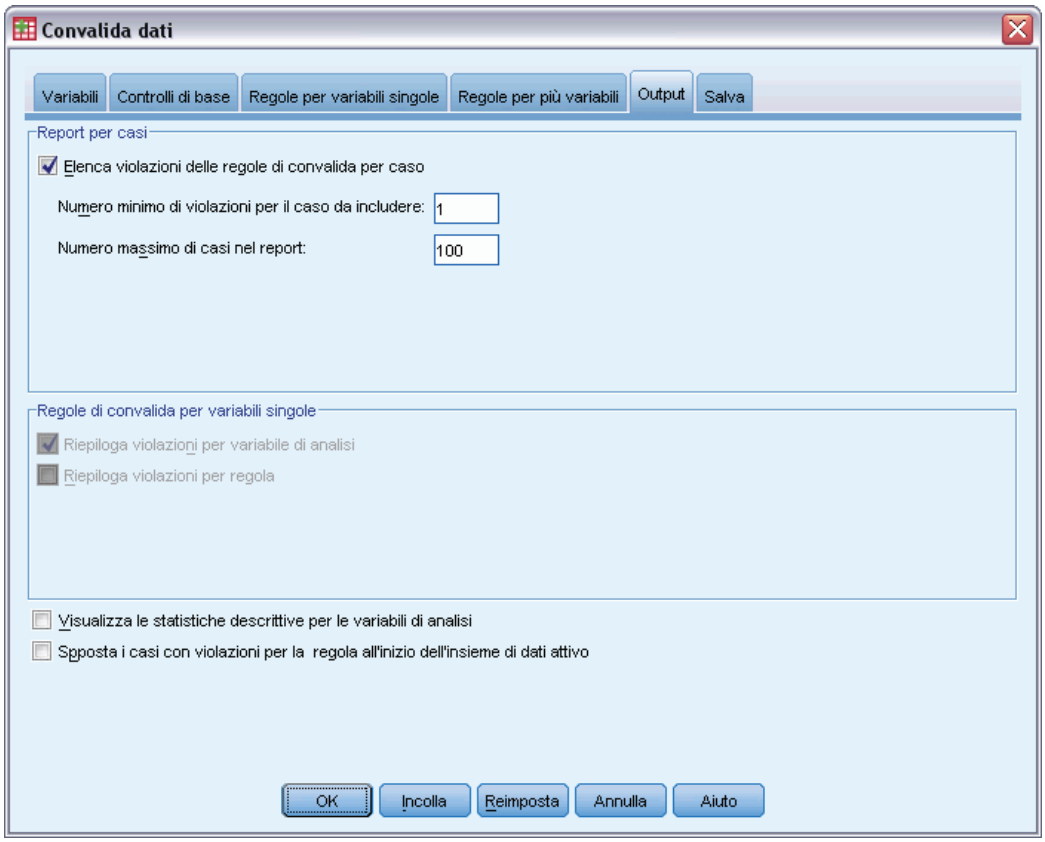
Figura 3-5
Scheda Regole per più variabili della finestra di dialogo Convalida dati



La scheda Regole per più variabili visualizza le regole di convalida disponibili per più variabili e consente di applicarle ai dati. Per definire ulteriori regole per più variabili, fare clic su Definisci regole. Per ulteriori informazioni, vedere l'argomento Definisci regole per più variabili in il capitolo 2 a pag. 6.

Output di Convalida dati

Figura 3-6
Scheda Output della finestra di dialogo Convalida dati



Report per casi. Se sono state applicate le regole per variabili singole o per più variabili, è possibile richiedere un report che elenchi le violazioni alle regole di convalida per ogni caso individuale.

- **Numero minimo di violazioni.** L'opzione specifica il numero minimo di violazioni alle regole richiesto per un caso, da includere nel report. Specificare un intero positivo.
- **Numero massimo di casi.** L'opzione specifica il numero massimo di casi inclusi nel report. Specificare un numero intero positivo, inferiore o uguale a 1000.

Regole di convalida per variabili singole. Se sono state applicate delle regole per variabili singole, è possibile scegliere come e se visualizzare i dati.

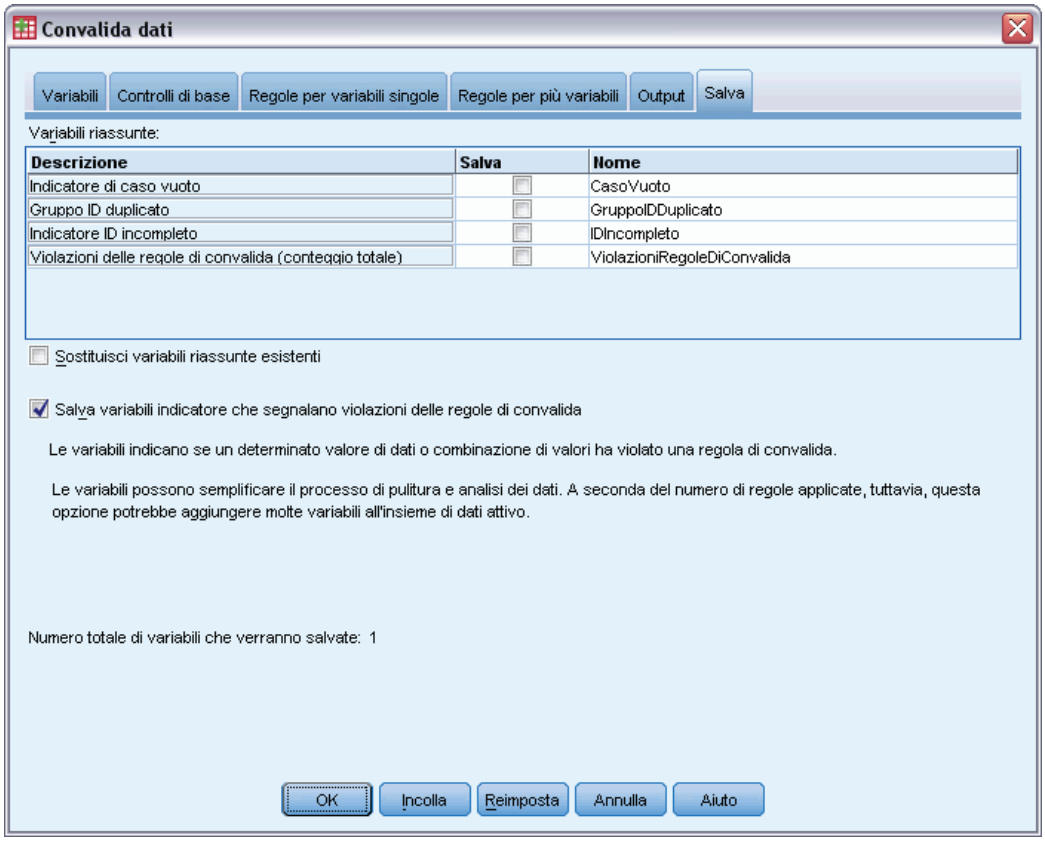
- **Riepiloga violazioni per variabile dell'analisi.** Per ogni variabile dell'analisi, l'opzione mostra le regole di convalida per variabili singole che sono state violate e il numero dei valori che ha violato ciascuna regola. L'opzione riporta inoltre il numero totale delle violazioni alle regole per variabili singole relativo a ogni variabile.
- **Riepiloga violazioni per regola.** Per ogni regola di convalida per variabili singole, l'opzione riporta le variabile che hanno violato la regola e il numero di valori non validi per ogni variabile. L'opzione riporta inoltre il numero totale dei valori che hanno violato ogni regola nelle variabili.

Visualizza le statistiche descrittive. L'opzione consente di richiedere delle statistiche descrittive per le variabili dell'analisi. Viene visualizzata una tabella di frequenza per ogni variabile categoriale. Per le variabili di scala viene generata una tabella delle statistiche riassuntive che include la media, la deviazione standard, il minimo e il massimo.

Sposta casi con violazioni delle regole di convalida. L'opzione consente di spostare i casi con violazioni alle regole per variabili singole e per più variabili all'inizio del file di dati attivo, per una lettura facile e accurata.

Salvataggio di Convalida dati

Figura 3-7
Scheda Salva della finestra di dialogo Convalida dati



La scheda Salva consente di salvare nel file di dati attivo le variabili che registrano delle violazioni alle regole.

Variabili riassuntive. Si tratta di variabili individuali che è possibile salvare. Contrassegnare una casella per salvare la variabile. Vengono forniti i nomi predefiniti per le variabili ed 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 nella scheda Controlli di base.

- **Gruppo ID duplicato.** Ai casi che presentano lo stesso identificatore di casi (tranne i casi con identificatori incompleti) viene assegnato lo stesso numero di gruppo. I casi con identificatori unici o incompleti vengono codificati 0.
- **Indicatore ID incompleto.** Ai casi con identificatori di casi vuoti o incompleti viene assegnato il valore 1. Tutti gli altri casi vengono codificati 0.
- **Violazioni delle regole di convalida.** Si tratta del conteggio totale per casi delle violazioni delle regole di convalida per variabili singole e più variabili.

Sostituisci variabili riassunte esistenti. Il nome delle variabili salvate nel file di dati deve essere unico, altrimenti sostituire le variabili con lo stesso nome.

Salva variabili indicatore. L'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 presenta il valore 1 se il caso viola la regola, se invece non la viola, ha valore 0.

Capitolo
4

Preparazione automatica dati

La preparazione dei dati per l'analisi è una delle fasi più importanti in qualsiasi progetto e, in genere, una delle più lunghe. 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. L'algoritmo si può utilizzare in modo completamente **automatico**, permettendogli di scegliere e applicare le correzioni, oppure in modo **interattivo**, visualizzando in anteprima le modifiche prima che vengano apportate e accettandole o rifiutandole in funzione delle esigenze.

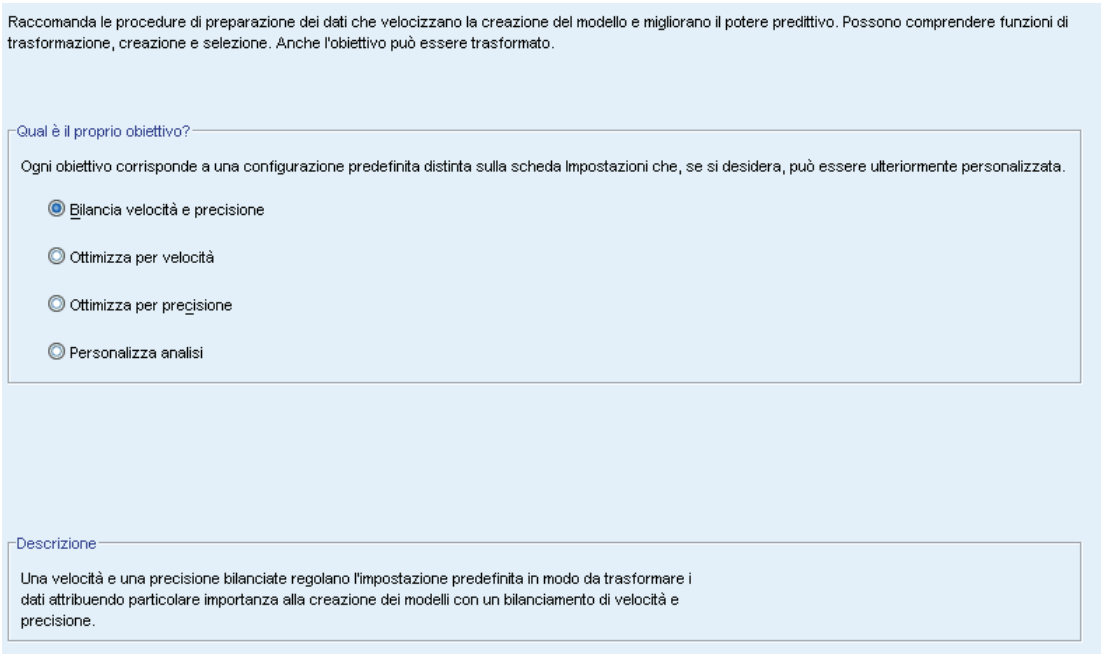
L'utilizzo di ADP consente di predisporre i dati per la creazione dei modelli in modo semplice e rapido, senza che sia necessario conoscere i concetti statistici impiegati. La creazione e il calcolo del punteggio dei modelli tenderanno a essere più rapidi; inoltre, l'utilizzo di ADP migliora la robustezza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

Nota: quando ADP prepara un campo per l'analisi, crea un nuovo campo che contiene le correzioni o le trasformazioni anziché sostituire i valori e le proprietà esistenti nel vecchio campo. Il vecchio campo non viene utilizzato nelle analisi successive e il suo ruolo viene impostato su Nessuno. Si noti inoltre che le informazioni sui valori mancanti definiti dall'utente non vengono trasferite nei nuovi campi e che tutti i valori mancanti nel nuovo campo sono valori 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. Per ulteriori informazioni, vedere l'argomento Utilizzo interattivo di Preparazione automatica dati in il capitolo 8 a pag. 84.

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. Per ulteriori informazioni, vedere l'argomento Utilizzo automatico di Preparazione automatica dati in il capitolo 8 a pag. 95.

Figura 4-1
Scheda Obiettivo di Preparazione automatica dati



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

- **Bilancia velocità e precisione.** 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 accedere alla Preparazione automatica dati

Dai menu, scegliere:

Trasforma > Prepara dati per la modellazione > Automatica...

▶ Fare clic su Esegui.

Se lo si desidera, è possibile:

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

Per accedere alla Preparazione interattiva dati

Dai menu, scegliere:
Trasforma > Prepara dati per la modellazione > Interattiva...

▶ Fare clic su Analizza nella barra degli strumenti nella parte superiore della finestra di dialogo.

▶ Fare clic sulla scheda Analisi ed esaminare le fasi di preparazione dei dati suggerite.

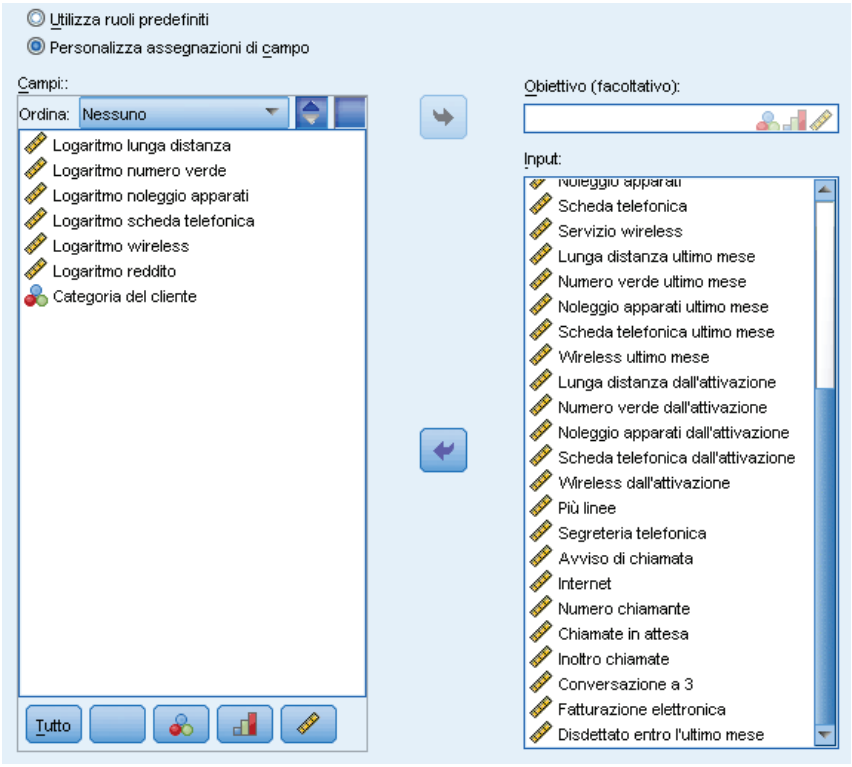
▶ Se sono corrette, fare clic su Esegui. In caso contrario, fare clic su Cancella analisi, modificare le impostazioni nel modo desiderato e selezionare Analizza.

Se lo si desidera, è possibile:

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

Scheda Campi

Figura 4-2
Scheda Campi di Preparazione automatica dati



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 obbligatorio avere 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:

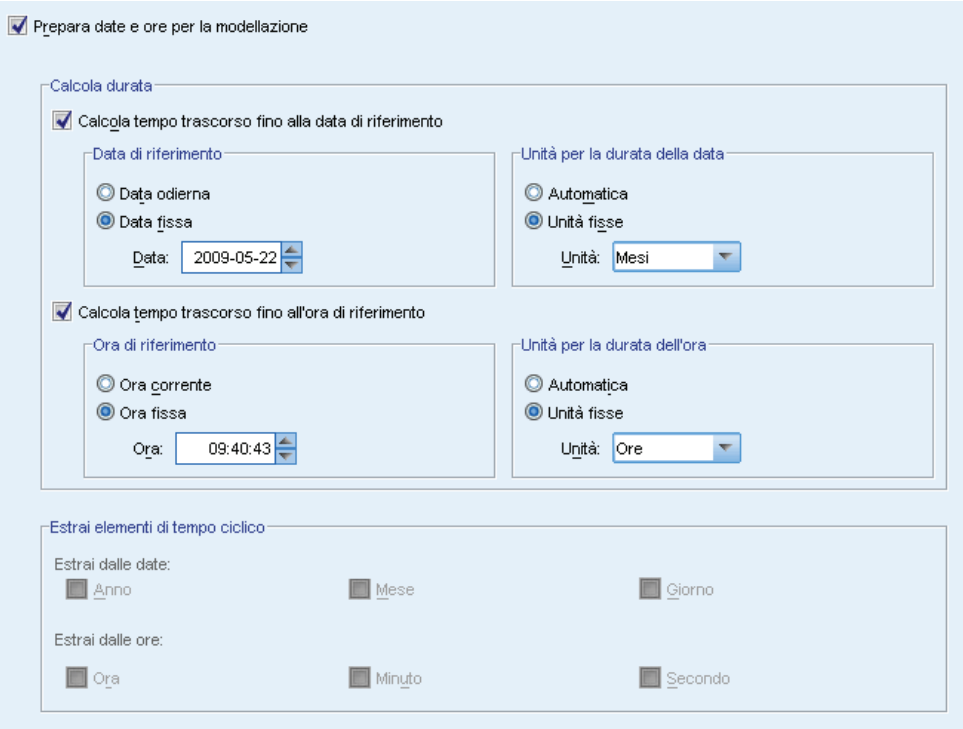
- **Obiettivo (facoltativo).** Se i modelli da creare richiedono un obiettivo, selezionare il campo obiettivo. Questa operazione equivale a impostare il ruolo del campo su Obiettivo.
- **Input.** Selezionare uno o più campi di input. Questa operazione equivale a impostare il ruolo del campo su Input.

Scheda Impostazioni

La scheda Impostazioni contiene vari gruppi di impostazioni che è possibile modificare per perfezionare l'elaborazione dei dati da parte dell'algoritmo. Se si apportano modifiche alle impostazioni predefinite che risultano incompatibili con gli altri obiettivi, la scheda Obiettivo viene aggiornata automaticamente per selezionare l'opzione Personalizza analisi.

Prepara date e ore

Figura 4-3
Preparazione automatica dati: impostazioni di Prepara date e ore



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

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

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

- **Data di riferimento.** Specifica la data a partire da cui sarà calcolata la durata relativamente alle informazioni sulla data presenti nei dati di input. Se si seleziona Data odierna, viene sempre utilizzata la data corrente del sistema per l'esecuzione di 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.

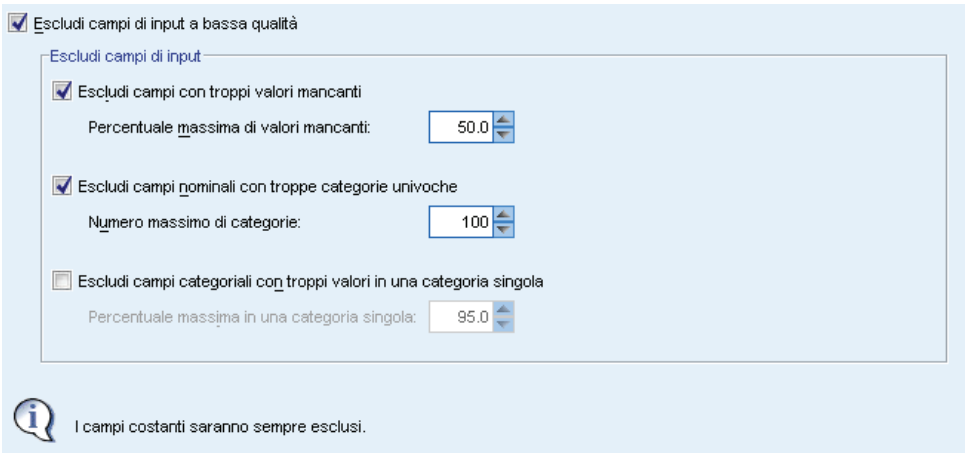
- **Ora di riferimento.** Specifica l'ora a partire dalla quale sarà calcolata la durata relativamente alle informazioni sull'ora presenti nei dati di input. Se si seleziona Ora corrente, viene sempre utilizzata l'ora corrente del sistema per l'esecuzione di ADP. Per utilizzare un'ora specifica, selezionare 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 si selezionano tutte e tre le caselle di controllo della data, il campo data di input "1954-05-23" viene suddiviso in tre campi: 1954, 5 e 23. Ciascuno utilizza il suffisso definito nel riquadro Nomi dei campi, e il campo 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

Figura 4-4
Preparazione automatica dati: impostazioni di 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 controlli Escludi campi ma vengono mantenute le selezioni.

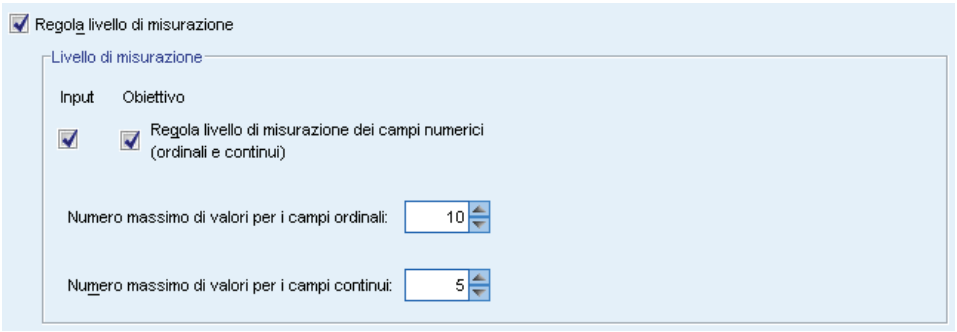
Escludi campi con troppi valori mancanti. I campi con una percentuale di valori mancanti superiore a quella specificata vengono eliminati dalla successiva analisi. Specificare un valore superiore o uguale a 0, che equivale a deselezionare questa opzione, e inferiore o uguale a 100, benché i campi con 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 esclusive del record provenienti dalla creazione del modello, quali ID, indirizzo o nome.

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

Regola misurazione

Figura 4-5
Preparazione automatica dati: impostazioni di Regola misurazione



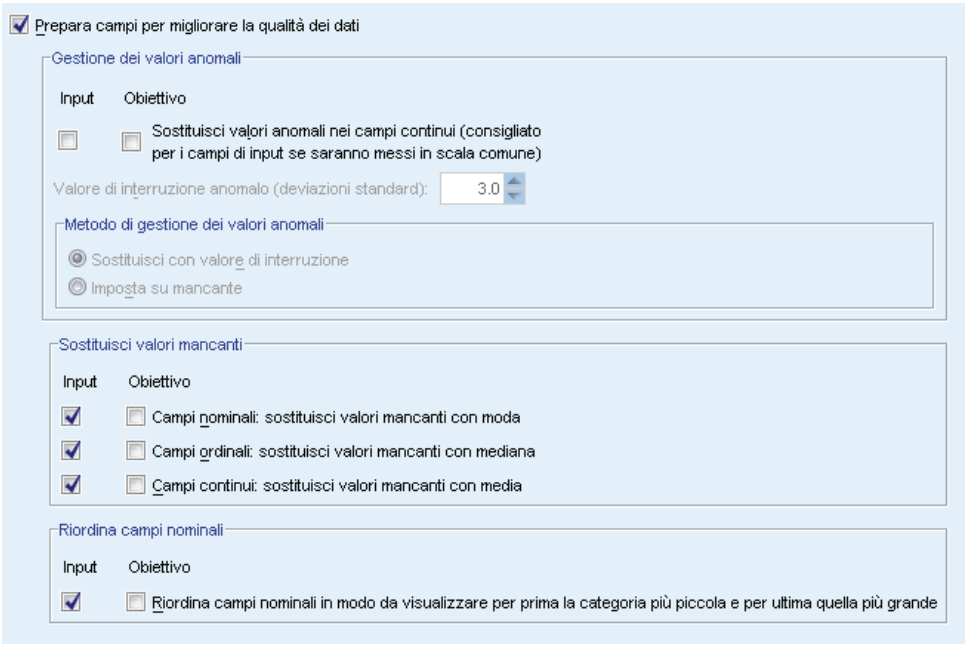
Regola livello di misurazione. Se si deseleziona questa opzione vengono disabilitati tutti gli altri controlli Regola misurazione ma vengono mantenute le selezioni.

Livello di misurazione. Specificare se il livello di misurazione dei campi continui con valori "insufficienti" può essere trasformato in ordinale e se i campi ordinali con valori "in eccesso" possono essere trasformati in continui.

- **Numero massimo di valori per i campi ordinali.** I campi ordinali con un numero di categorie superiore a quello specificato vengono riformulati come campi continui. Specificare un intero positivo. L'impostazione predefinita è 10. Questo valore deve essere superiore o uguale al numero minimo di valori per i campi continui.
- **Numero minimo di valori per i campi continui.** I campi continui con un numero di valori univoci inferiore a quello specificato vengono riformulati come campi ordinali. Specificare un intero positivo. L'impostazione predefinita è 5. Questo valore deve essere inferiore o uguale al numero massimo di valori per i campi ordinali.

Migliora qualità dei dati

Figura 4-6
Preparazione automatica dati: impostazioni di Migliora qualità dei dati



Prepara campi per migliorare la qualità dei dati. Se si deseleziona questa opzione vengono disabilitati tutti gli altri controlli Migliora qualità dei dati ma vengono mantenute le selezioni.

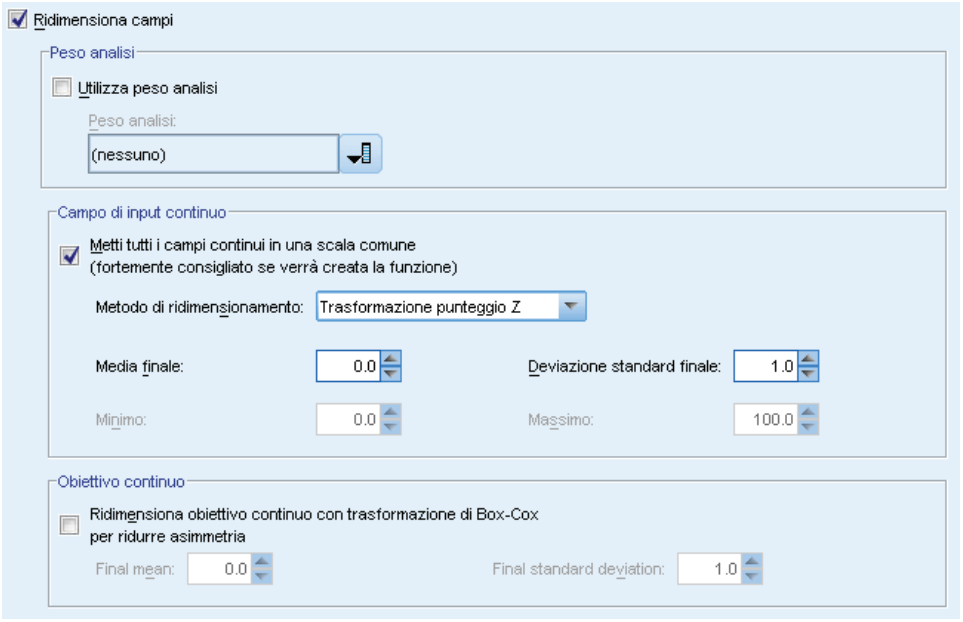
Gestione dei valori anomali. Specificare se i valori anomali degli input e dell'obiettivo devono essere sostituiti; in caso affermativo, specificare un punto di interruzione dei valori anomali, misurato in deviazioni standard, e un metodo di sostituzione dei valori anomali. I valori anomali si possono sostituire tagliandoli (impostandoli sul punto di interruzione) o impostandoli come valori mancanti. Tutti i valori anomali impostati su valori mancanti rispondono alle impostazioni per la gestione dei valori mancanti selezionate di seguito.

Sostituisci valori mancanti. Specificare se i valori mancanti dei campi continui, nominali o ordinali devono essere sostituiti.

Riordina campi nominali. Selezionare questa opzione per ricodificare i valori dei campi (insieme) nominali dalla categoria minima (che ricorre con minore frequenza) a quella massima (che ricorre con maggiore frequenza). I nuovi valori dei campi iniziano con 0 come categoria meno frequente. Si noti 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.

Ridimensiona campi

Figura 4-7
Preparazione automatica dati: impostazioni di Ridimensiona campi



Ridimensiona campi. Se si deseleziona questa opzione vengono disabilitati tutti gli altri controlli Ridimensiona campi ma vengono mantenute le selezioni.

Peso analisi. Questa variabile contiene i pesi delle analisi (regressione o campionamento). I pesi delle analisi si utilizzano per tenere conto delle differenze nella varianza tra i vari livelli del campo obiettivo. Seleziona un campo continuo.

Campi di input continui. Questa opzione normalizza i campi di input continui utilizzando una trasformazione punteggio Z o una trasformazione Min/Max. Il ridimensionamento degli input è utile soprattutto quando si seleziona Esegui creazione funzioni nelle impostazioni Seleziona e crea.

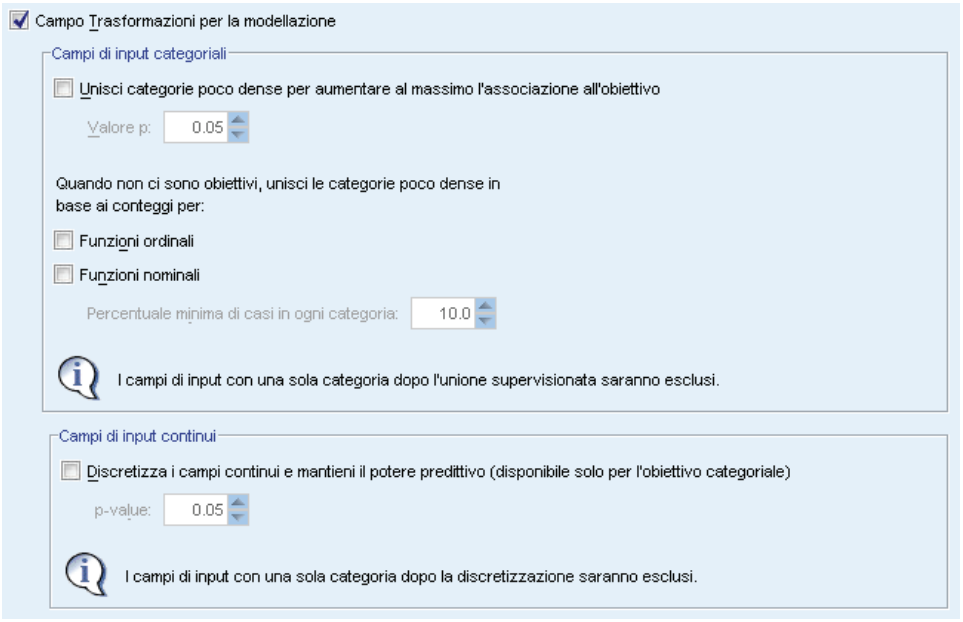
- **Trasformazione punteggio Z.** Utilizzando la media osservata e la deviazione standard come stime dei parametri relativi alla popolazione, i campi vengono standardizzati e quindi i punteggi *Z* vengono associati ai valori corrispondenti di una distribuzione normale con la Media finale e la Deviazione standard finale specificate. Indicare un numero per la Media finale e un numero positivo per la Deviazione standard finale. I valori predefiniti sono rispettivamente 0 e 1, corrispondenti alla modifica della scala standardizzata.
- **Trasformazione Min/Max.** Utilizzando il minimo e il massimo osservati come stime dei parametri relativi alla popolazione, i campi vengono associati ai valori corrispondenti di una distribuzione uniforme con il Minimo e il Massimo specificati. Specificare i numeri con un Massimo superiore al Minimo.

Obiettivo continuo. Mediante la trasformazione di Box-Cox, questa opzione trasforma un obiettivo continuo in un campo con una distribuzione più o meno normale con la Media finale e la Deviazione standard finale specificate. Indicare un numero per la Media finale e un numero positivo per la Deviazione standard finale. I valori predefiniti sono rispettivamente 0 e 1.

Nota: Se un obiettivo è stato trasformato da ADP, i modelli successivi creati con l'obiettivo trasformato calcolano il punteggio delle unità trasformate. Per interpretare e utilizzare questi risultati è necessario riconvertire il valore previsto nella scala originale. Per ulteriori informazioni, vedere l'argomento Trasforma all'indietro i punteggi a pag. 45.

Trasforma campi

Figura 4-8
Preparazione automatica dati: impostazioni di Trasforma campi



Per migliorare il potere predittivo dei dati è possibile trasformare i campi di input.

Campo Trasformazioni per la modellazione. Se si deseleziona questa opzione vengono disabilitati tutti gli altri controlli Trasforma campi ma vengono mantenute le selezioni.

Campi di input categoriali

- **Unisci categorie poco dense 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 presentano differenze significative (ovvero che hanno un valore *p* superiore al valore specificato) vengono unite. Specificare un valore superiore a 0 e inferiore o uguale a 1. Se tutte le categorie vengono unite in una sola, le versioni originali e derivate del campo vengono escluse da ulteriori analisi perché prive di valore come predittori.
- **Quando non ci sono obiettivi, unisci le categorie poco dense in base ai conteggi.** Se l'insieme di dati non ha un obiettivo, è possibile decidere di unire le categorie poco dense di campi ordinali e nominali. Per unire le categorie con meno della percentuale minima specificata del numero totale di record viene utilizzato il metodo delle frequenze uguali. Specificare un valore superiore o uguale a 0 e inferiore o uguale a 100. L'impostazione predefinita è 10.

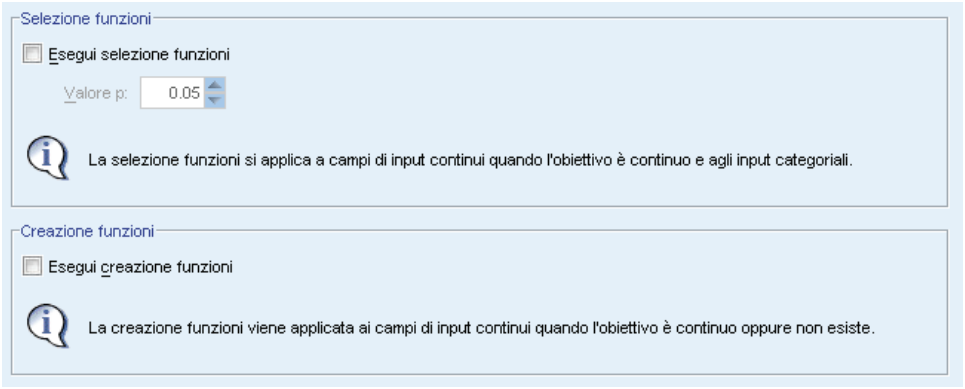
L'unione si interrompe quando non ci sono più categorie con meno della percentuale minima specificata di casi, o quando rimangono solo due categorie.

Campi di input continui. Se l'insieme di dati comprende un obiettivo categoriale, è possibile categorizzare gli input continui con associazioni forti per migliorare le prestazioni in sede di elaborazione. Gli intervalli vengono creati in base alle proprietà di "sottoinsiemi omogenei", che vengono identificati mediante il metodo di Scheffe utilizzando il valore *p* specificato come alfa per il valore critico per la determinazione degli insiemi omogenei. Specificare un valore maggiore di 0 e minore o uguale a 1. L'impostazione predefinita 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: la categorizzazione in ADP è diversa dalla categorizzazione ottimale. La categorizzazione ottimale utilizza le informazioni relative all'entropia per convertire un campo continuo in uno categoriale; per far questo è necessario ordinare i dati e archiviarli tutti in memoria. ADP utilizza sottoinsiemi omogenei per categorizzare un campo continuo e, pertanto, la categorizzazione in ADP non richiede di ordinare i dati e non li archivia tutti in memoria. Quando si utilizza il metodo dei sottoinsiemi omogenei per categorizzare un campo continuo, il numero di categorie dopo la categorizzazione è sempre inferiore o uguale al numero di categorie dell'obiettivo.

Seleziona e crea

Figura 4-9
Preparazione automatica dati: impostazioni di Seleziona e crea



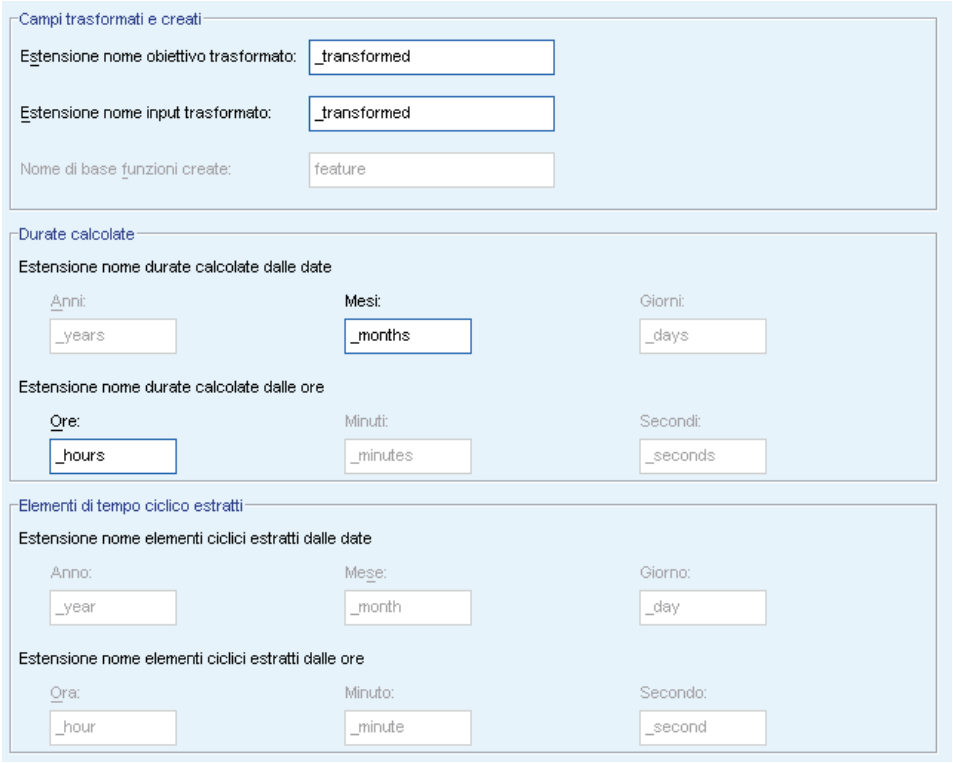
Per migliorare il potere predittivo dei dati è possibile creare nuovi campi basati su quelli esistenti.

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

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

Nomi campi

Figura 4-10
Preparazione automatica dati: impostazioni di Nomina 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 saranno denominate: da funzione1 a funzione9.
- le funzioni create da 10 a 99 saranno denominate: da funzione01 a funzione99.
- le funzioni create da 100 a 999 saranno denominate: da funzione001 a funzione999, 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.

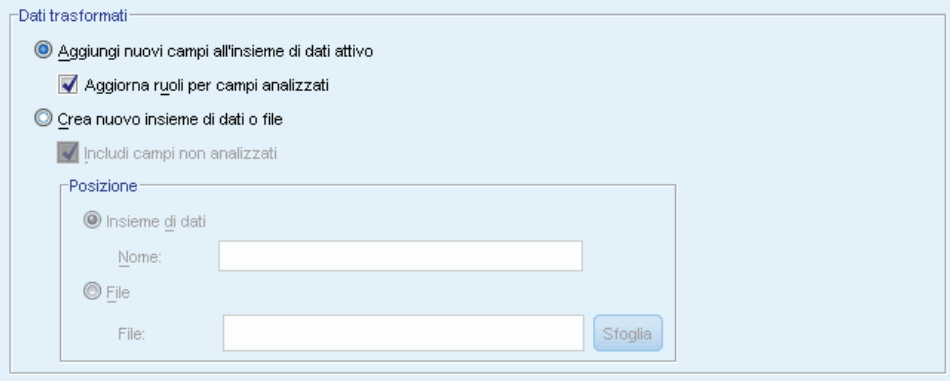
Elementi ciclici estratti da date e ore. Specificare le estensioni dei nomi da applicare agli elementi ciclici estratti da date e ore.

Applicazione e salvataggio delle trasformazioni

A seconda che si utilizzi la finestra di dialogo Preparazione automatica dati o Preparazione interattiva dati, le impostazioni per applicare e salvare le trasformazioni sono leggermente diverse.

Preparazione interattiva dati: impostazioni di Applica trasformazioni

Figura 4-11
Preparazione interattiva dati: impostazioni di Applica trasformazioni

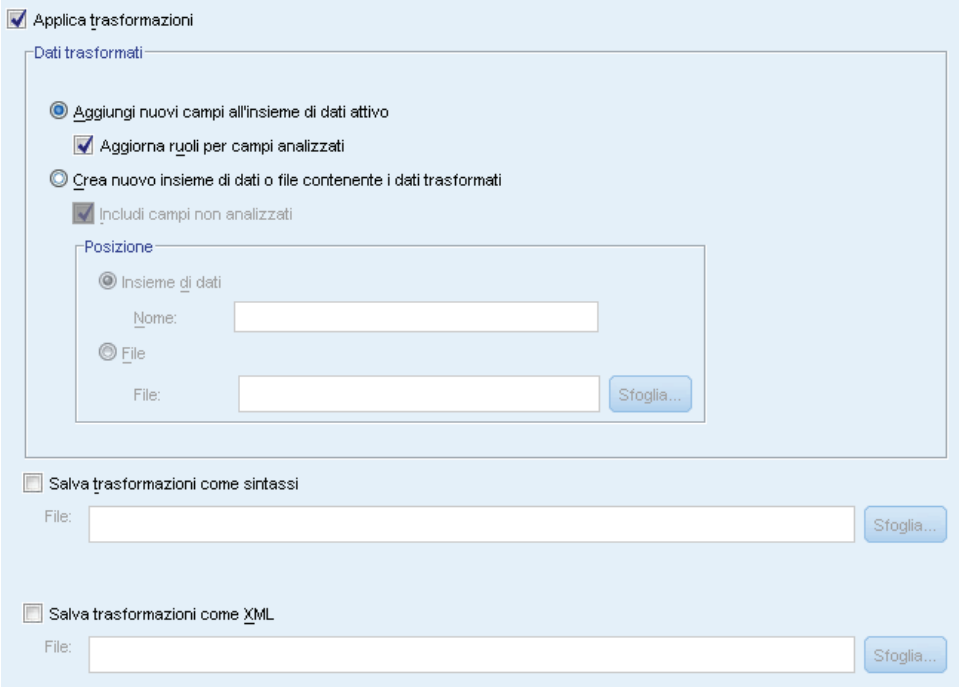


Dati trasformati. Queste impostazioni definiscono dove salvare i dati trasformati.

- **Aggiungi nuovi campi all'insieme di dati attivo.** Tutti i campi creati dalla preparazione automatica dati vengono aggiunti come nuovi campi all'insieme di dati attivo. Aggiorna ruoli per campi analizzati imposta il ruolo su Nessuno per tutti i campi esclusi da ulteriori analisi dalla preparazione automatica dati.
- **Crea nuovo insieme di dati o file contenente i dati trasformati.** I campi consigliati dalla preparazione automatica dati vengono aggiunti a un nuovo insieme di dati o file. Includi campi non analizzati aggiunge al nuovo insieme di dati i campi dell'insieme di dati originale che non erano stati specificati nella scheda Campi. Si tratta di un'opzione utile per trasferire nel nuovo insieme di dati i campi che contengono informazioni non utilizzate nella creazione del modello, quale l'ID, l'indirizzo o il nome.

Preparazione automatica dati: impostazioni di Applica e salva

Figura 4-12
Preparazione automatica dati: impostazioni di Applica e salva



Il gruppo Dati trasformati è lo stesso presente in Preparazione interattiva dati. In Preparazione automatica dati sono disponibili le seguenti opzioni aggiuntive:

Applica trasformazioni. Nelle finestre di dialogo di Preparazione automatica dati, se si deseleziona questa opzione si deselezionano tutti gli altri controlli Applica e Salva ma vengono mantenute le selezioni.

Salva trasformazioni come sintassi. Questa opzione salva le trasformazioni consigliate come sintassi dei comandi in un file esterno. La finestra di dialogo Preparazione interattiva dati non dispone di questo comando perché incolla le trasformazioni come sintassi dei comandi nella finestra della sintassi se si fa clic su Incolla.

Salva trasformazioni come XML. Questa opzione salva le trasformazioni consigliate come XML in un file esterno, che è possibile unire al file PMML del modello mediante il comando `TMS MERGE` o applicare a un altro insieme di dati mediante il comando `TMS IMPORT`. La finestra di dialogo Preparazione interattiva dati non dispone di questo comando poiché salva le trasformazioni in formato XML se si fa clic su Salva XML nella barra degli strumenti nella parte superiore della finestra.

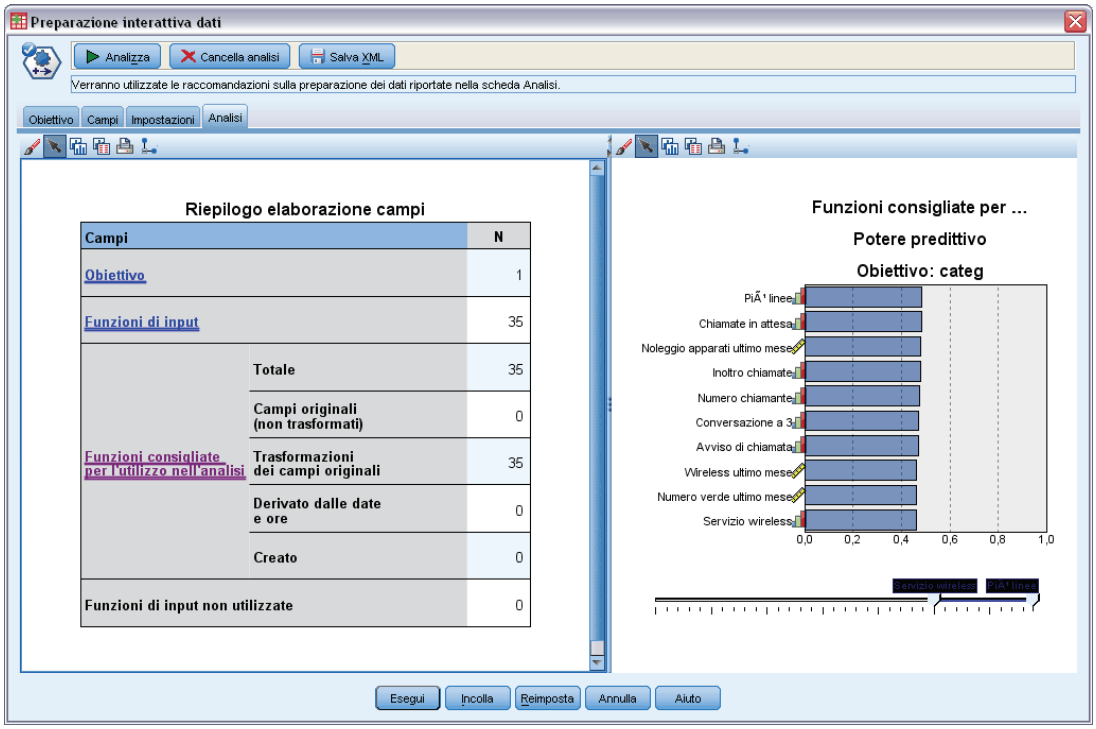
Scheda Analisi

Nota: la scheda Analisi della finestra di dialogo Preparazione interattiva dati consente di esaminare le trasformazioni consigliate. La finestra di dialogo Preparazione automatica dati non prevede questo passaggio.

- 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.

Figura 4-13
Scheda Analisi di Preparazione automatica dati



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, vedere l'argomento Riepilogo elaborazione campi a pag. 33.
- Campi. Per ulteriori informazioni, vedere l'argomento Campi a pag. 35.
- Riepilogo delle azioni Per ulteriori informazioni, vedere l'argomento Riepilogo delle azioni a pag. 37.

Le visualizzazioni collegate/ausiliarie sono quattro:

- Potere predittivo (impostazione predefinita). Per ulteriori informazioni, vedere l'argomento Potere predittivo a pag. 38.
- Tabella campi. Per ulteriori informazioni, vedere l'argomento Tabella Campi a pag. 39.
- Dettagli campo. Per ulteriori informazioni, vedere l'argomento Dettagli campo a pag. 40.
- Dettagli dell'azione. Per ulteriori informazioni, vedere l'argomento Dettagli dell'azione a pag. 42.

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

Figura 4-14
Riepilogo elaborazione campi

Riepilogo elaborazione campi

Campi		C
Obiettivo		1
Funzioni di input		9
Funzioni consigliate per l'utilizzo nell'analisi	Totale	7
	Campi originali (non trasformati)	2
	Trasformazioni dei campi originali	5
	Derivato dalle date e ore	0
	Creato	0
Funzioni di input non utilizzate		2

La tabella Riepilogo elaborazione campi offre un'istantanea 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.

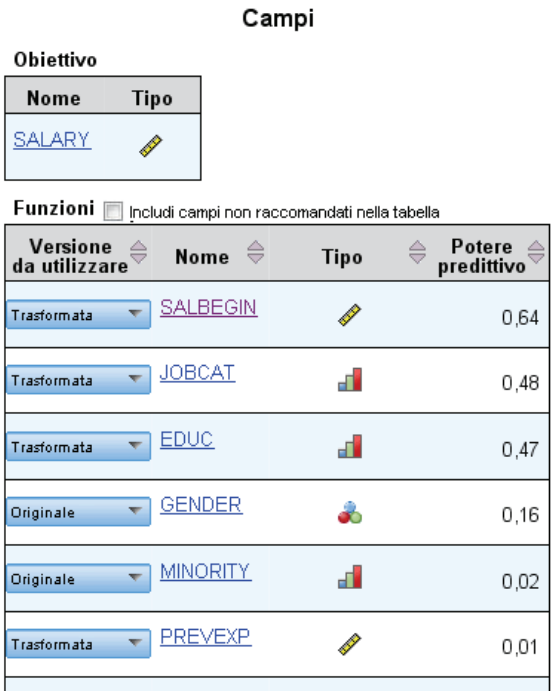
La tabella riporta le seguenti informazioni:

- Il numero di campi obiettivo.
- Il numero di predittori (input) originali.
- I predittori consigliati per l'uso nell'analisi e nella modellazione. Sono compresi il numero totale dei campi consigliato; il numero dei campi originali non trasformati consigliato; il numero dei campi trasformati consigliato (escludendo le versioni intermedie di qualsiasi campo, i campi derivati da predittori data/ora e i predittori creati); il numero consigliato dei campi derivati da campi data/ora e il numero consigliato dei predittori creati.
- 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, vedere l'argomento Tabella Campi a pag. 39. Le funzioni consigliate per l'analisi sono visualizzate nella visualizzazione collegata Potere predittivo. Per ulteriori informazioni, vedere l'argomento Potere predittivo a pag. 38.

Campi

Figura 4-15
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 (continua, 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 riporta tre opzioni: Trasformata, Originale e Non utilizzare.

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

  Per i campi derivati data/ora e i predittori creati, le opzioni 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 che hanno una versione originale e una trasformata, il passaggio dalla versione Originale a quella Trasformata aggiorna automaticamente le impostazioni Livello di misurazione e Potere predittivo.

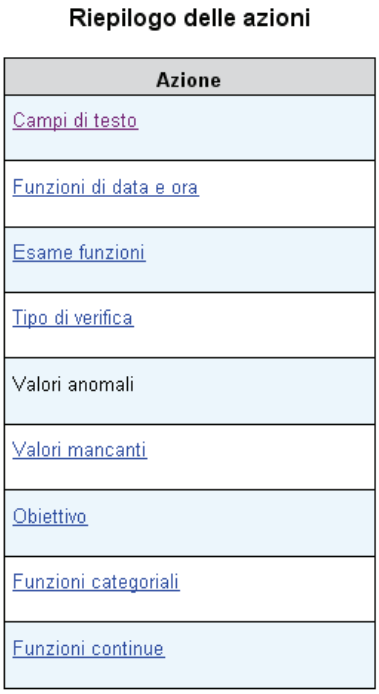
- **Nome.** Ogni nome di campo è un collegamento. Fare clic su un nome per visualizzare ulteriori informazioni sul campo nella visualizzazione collegata. Per ulteriori informazioni, vedere l'argomento Dettagli campo a pag. 40.

- **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 (continua, 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

Figura 4-16
Riepilogo delle azioni



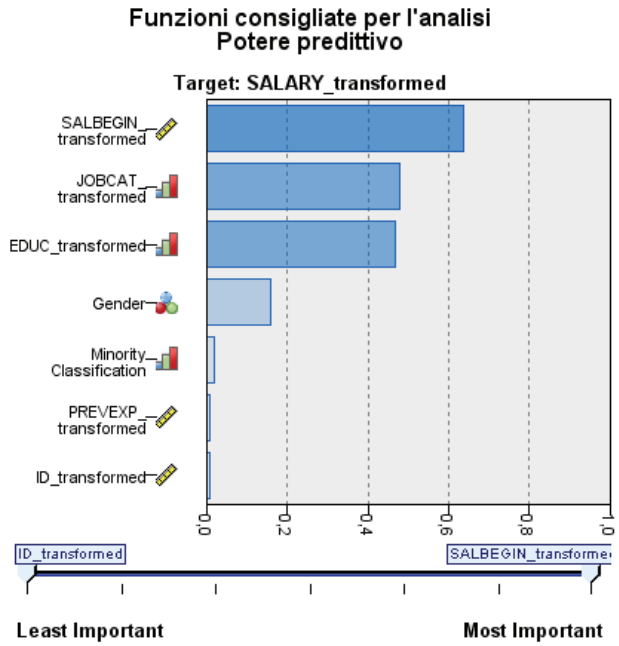
Per ogni azione svolta dalla preparazione dati automatica, i predittori di input vengono trasformati e/o eliminati; i campi che sopravvivono a un'azione vengono utilizzati nella successiva. I campi che superano tutti i passaggi sono quelli di cui si consiglia l'utilizzo nella modellazione, mentre gli 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 una Azione sottolineata, farvi clic sopra per visualizzare ulteriori dettagli sulle azioni intraprese in una visualizzazione collegata. Per ulteriori informazioni, vedere l'argomento Dettagli dell'azione a pag. 42.

Nota: sono visualizzate solo le versioni originali e trasformate definitive di ogni campo, non quelle intermedie utilizzate durante l'analisi.

Potere predittivo

Figura 4-17
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 rispecchia il suffisso scelto nel riquadro Nomi dei campi della scheda Impostazioni; ad esempio: *_transformed*.

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

Il potere predittivo di ogni predittore consigliato è calcolato da un modello di regressione lineare o naïve Bayes, a seconda che l'obiettivo sia continuo o categoriale.

Tabella Campi

Figura 4-18
Tabella campi

Funzioni di input

Nome	Tipo
ID	Continuo
GENDER	Insieme
BDATE	Continuo
EDUC	Insieme ordinato
JOBCAT	Insieme ordinato
SALBEGIN	Continuo
JOBTIME	Continuo
PREVEXP	Continuo
MINORITY	Insieme ordinato

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.** 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 rispecchia il suffisso scelto nel riquadro Nomi dei campi della scheda Impostazioni; ad esempio: *_transformed*.

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

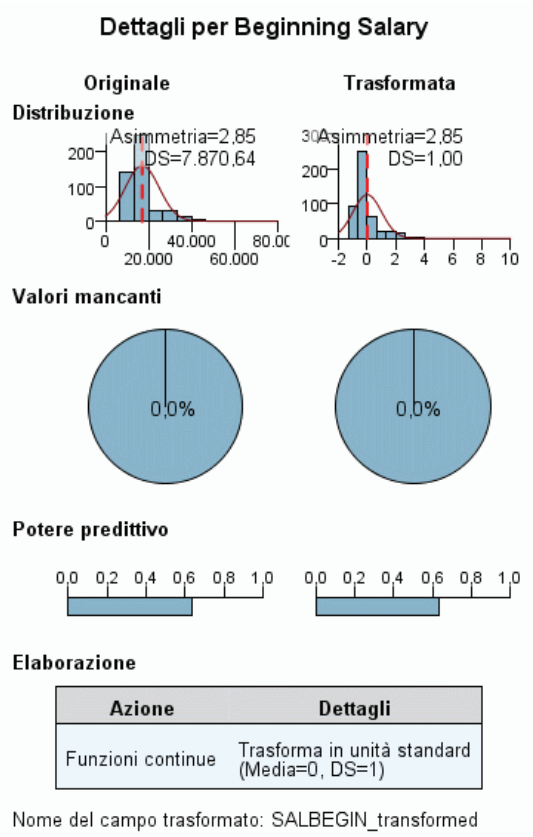
  Per i predittori creati viene utilizzato il nome del predittore creato, ad esempio: *Predittore1*.

- **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

Figura 4-19
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. Sono inoltre visualizzati la cronologia di elaborazione del campo e 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 ha 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 sono visualizzati se non è definito un obiettivo o se si fa clic sull'obiettivo nel riquadro 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 è 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, vedere l'argomento Dettagli dell'azione a pag. 42.
- **Dettagli.** L'elenco delle procedure eseguite. Ad esempio, Trasforma in unità standard.
- **Funzione.** Visualizzata solo per i predittori creati, mostra la combinazione lineare dei campi di input, ad esempio 0,06*age + 1,21*height.

Dettagli dell'azione

Figura 4-20
Analisi ADP - Dettagli dell'azione

Passaggio 9: Funzioni continue

Trasformazione	Numero di funzioni	Criteri	
		Media	DS
Trasforma in unità standard	5	0	1

Creazione spazio di funzioni	C
Funzioni create	0
Funzioni escluse a causa di una scarsa associazione all'obiettivo	2
Funzioni escluse perché costanti dopo la discretizzazione	0

Visualizzata quando si seleziona una Azione sottolineata nella visualizzazione principale Riepilogo delle azioni, la visualizzazione collegata Dettagli dell'azione mostra informazioni generali e specifiche di un'azione per ogni fase di elaborazione effettuata; i dettagli relativi alle singole azioni 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 un obiettivo, quando si fa clic sull'azione nella visualizzazione principale Riepilogo delle azioni viene visualizzato un messaggio di errore al posto dei dettagli dell'azione.

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

Tabella Campi di 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 Verifica 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 nessuno dei campi di input (obiettivo o predittori) era continuo o ordinale, questo è segnalato in una nota a piè di pagina.

Tabella Valori anomali

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

- Indica o il numero di campi continui per i quali sono stati trovati e tagliati dei valori anomali, oppure il numero di campi continui per i quali sono stati trovati e impostati come mancanti dei valori anomali, a seconda delle impostazioni nel riquadro 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 predittori) 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 poco dense.
- Escluse a causa di una scarsa associazione all'obiettivo.
- Escluse perché erano costanti dopo l'unione.

In assenza di 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 predittore 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:

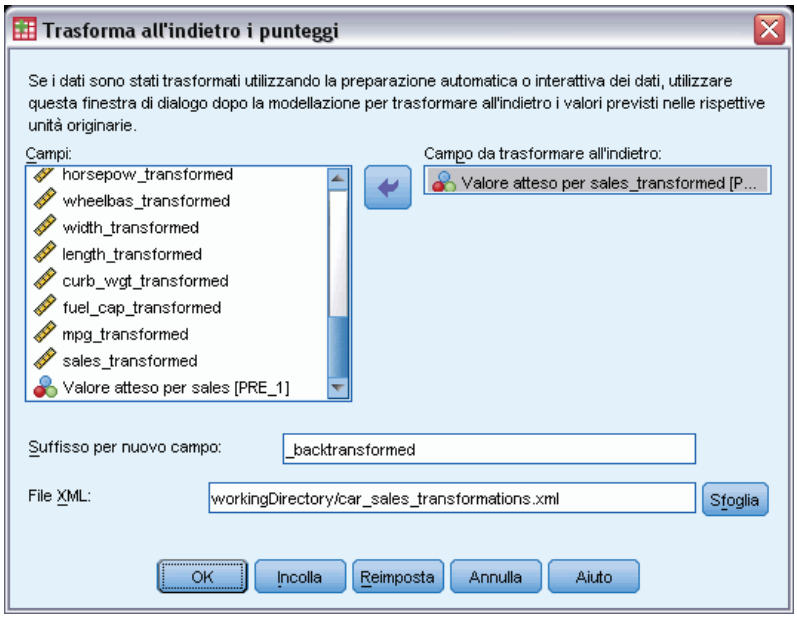
- Create.
- Escluse a causa di una scarsa associazione all'obiettivo.
- Escluse perché erano costanti dopo la categorizzazione.
- Escluse perché erano costanti dopo la creazione.

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 creati con l'obiettivo trasformato calcolano il punteggio delle unità trasformate. Per interpretare e utilizzare questi risultati è necessario riconvertire il valore previsto nella scala originale.

Figura 4-21
Trasforma all'indietro i punteggi



Per trasformare all'indietro i punteggi, dai menu scegliere:
Trasforma > Prepara dati per la modellazione > Trasforma all'indietro i punteggi...

- Selezionare un campo da trasformare all'indietro. Il campo deve contenere i valori previsti dal modello dell'obiettivo trasformato.
- Specificare un suffisso per il nuovo campo. Il nuovo campo conterrà i valori previsti dal modello nella scala originale dell'obiettivo non trasformato.
- Specificare il percorso del file XML contenente le trasformazioni ADP. Il file deve essere stato salvato a partire dalle finestre di dialogo Preparazione interattiva dati o Preparazione automatica dati. Per ulteriori informazioni, vedere l'argomento Applicazione e salvataggio delle trasformazioni a pag. 30.

Capitolo
5

Identifica casi anomali

La procedura Rilevamento anomalie ricerca i casi insoliti in base agli scostamenti dalle norme dei gruppi cluster corrispondenti. La procedura permette di rilevare rapidamente i casi insoliti per il controllo dei dati nella fase di analisi esplorativa dei dati, prima di effettuare l'analisi inferenziale dei dati. Questo algoritmo è progettato per il rilevamento di casi anomali generici. In altre parole la definizione di casi anomali non si riferisce a un'applicazione specifica, come il rilevamento degli schemi di pagamento anomali nell'industria sanitaria o il rilevamento di casi di riciclaggio nell'industria finanziaria in cui un'anomalia può essere definita in modo specifico.

Esempio. Ad un analista viene chiesto di preparare dei modelli predittivi per valutare i risultati dei trattamenti ai pazienti vittime di infarti cardiaci. L'analista è preoccupato della qualità dei dati perché questi modelli sono sensibili alle osservazioni anomale. Alcune di queste osservazioni anomale rappresentano di fatto casi univoci e non sono indicati per la previsione, mentre altre sono dovute ad errori di inserimento dati, ovvero a casi in cui i valori sono tecnicamente "corretti" che non vengono quindi rilevati dalle procedure di convalida dati. La procedura Identifica casi anomali ricerca e trova tutti i valori anomali permettendo all'analista di decidere come gestirli.

Statistiche. Questa procedura genera gruppi equivalenti, norme di gruppi equivalenti per le variabili continue e categoriali, indici di anomalie basate sugli scostamenti rispetto alle norme dei gruppi equivalenti e i valori di impatto delle variabili che nella maggior parte dei casi determinano la condizione di anomalia.

Considerazioni sui dati

Dati. Questa procedura può essere utilizzata sia con le variabili continue sia con le variabili categoriali. Ciascuna riga rappresenta un'osservazione specifica e ciascuna colonna rappresenta la variabile specifica su cui sono basati i gruppi analoghi. Il file di dati può contenere una variabile di casi per l'identificazione che può essere utilizzata per contrassegnare l'output, ma non per l'analisi. L'uso di valori mancanti è consentito. La variabile di peso, se specificata, viene ignorata.

Il modello di rilevamento può essere applicato a un nuovo file dati di prova. Gli elementi dei dati di prova devono essere uguali agli elementi dei dati utilizzati per la formazione. E, a seconda delle impostazioni dell'algoritmo, la modalità di gestione dei dati mancanti usata per creare il modello può anche essere applicata al file dei dati di prova prima del calcolo del punteggio.

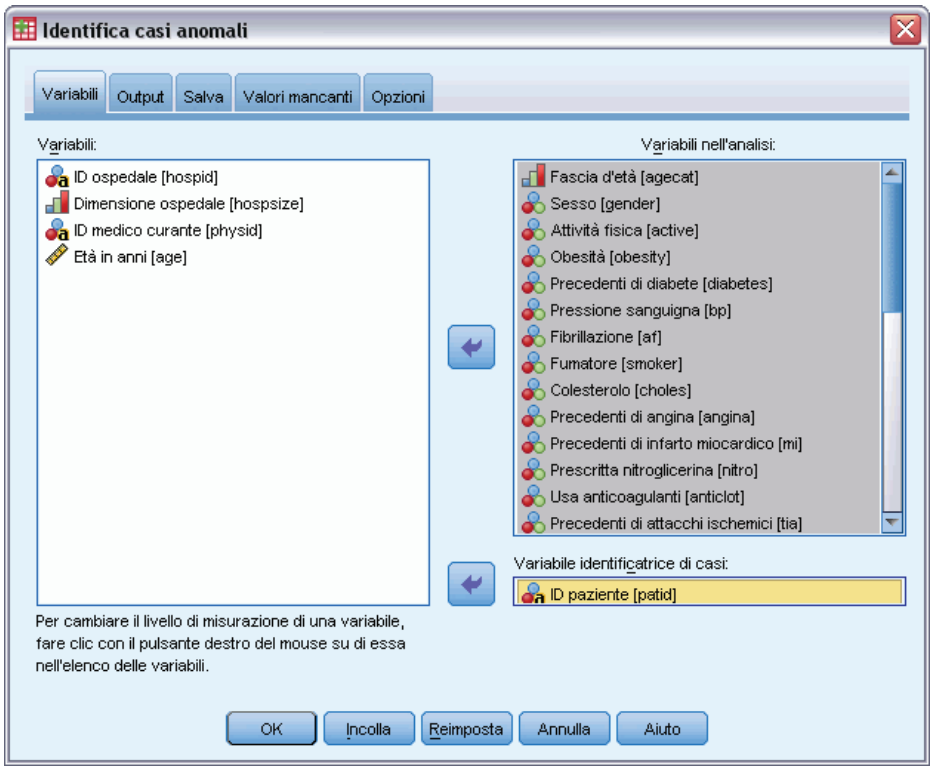
Ordine dei casi. È utile notare che la soluzione può dipendere dall'ordine dei casi. Per ridurre al minimo gli effetti dell'ordine, disporre i casi in ordine casuale. Per verificare la stabilità di una data soluzione, può essere utile ottenere più soluzioni diverse con casi disposti in ordini casuali diversi. Se i file sono particolarmente grandi, è possibile eseguire più analisi con un campione di casi disposti in più ordini casuali.

Assunzioni. L'algoritmo presume che tutte le variabili siano non costanti e indipendenti nonché che i casi non abbiano valori mancanti per le variabili di input. Si suppone che a ogni variabile continua sia associata una distribuzione normale o gaussiana, mentre a ogni variabile categoriale una distribuzione multinomiale. La verifica empirica interna indica che la procedura è abbastanza resistente alle violazioni delle assunzioni di indipendenza e distribuzione. Tuttavia, è necessario verificare fino a che punto tali assunzioni vengono soddisfatte.

Per identificare i casi anomali

- Dai menu, scegliere:
  Dati > Identifica casi anomali...

Figura 5-1
Scheda Variabili della finestra di dialogo Identifica casi anomali

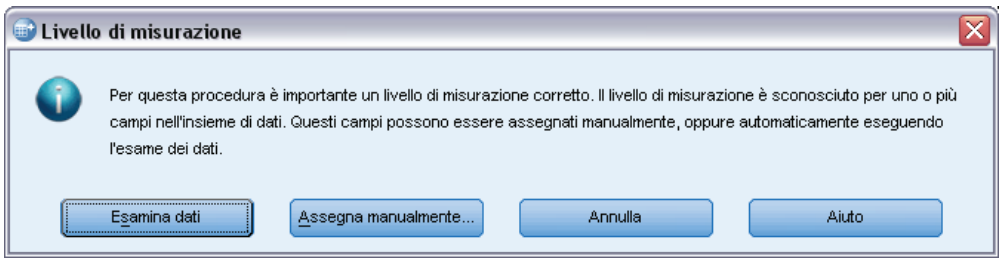


- Selezionare almeno una variabile di analisi.
- Oppure scegliere una variabile di identificazione dei casi per usare l'output delle etichette.

Campi con livello di misurazione sconosciuto

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

Figura 5-2
Avviso Livello di misurazione

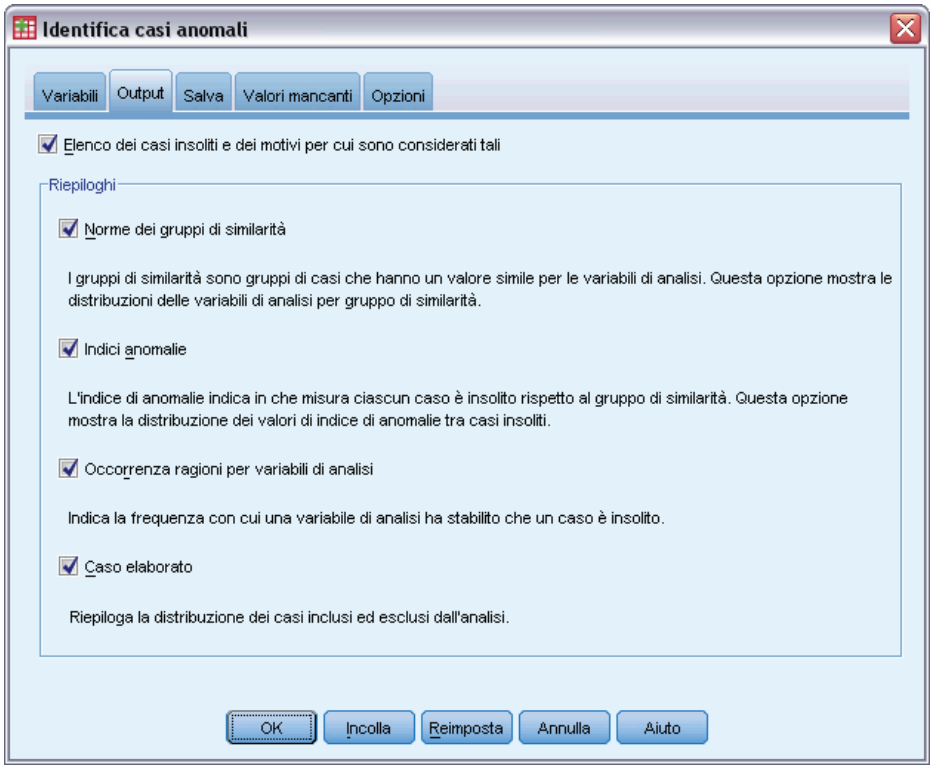


- **Esamina dati.** Legge i dati dell'insieme di dati attivo e assegna un livello di misurazione predefinito a tutti i campi con livello di misurazione sconosciuto. Con insiemi di dati di grandi dimensioni, questa operazione può richiedere del tempo.
- **Assegna manualmente.** Apre una finestra di dialogo 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 Visualizzazione variabili 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.

Identifica l'output di casi anomali

Figura 5-3
Scheda Output della finestra di dialogo Identifica casi anomali



Elenca i casi anomali e i motivi per cui vengono considerati tali. Questa opzione rende disponibili tre tabelle:

- L'elenco Indice dei casi anomali visualizza i casi identificati come anomali unitamente all'indice di anomalia corrispondente.
- L'elenco ID casi anomali equivalenti visualizza i casi anomali e le informazioni relativi ai gruppi equivalenti corrispondenti.
- L'elenco Motivi anomalie visualizza il numero di caso, la variabile del motivo, il valore di impatto della variabile, il valore della variabile e la norma della variabile per ciascun motivo.

Tutte le tabelle sono disposte in ordine decrescente in base all'indice di anomalia. Inoltre, vengono anche visualizzati gli ID dei casi se la variabile di identificazione del caso è stata specificata nella scheda Variabili.

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

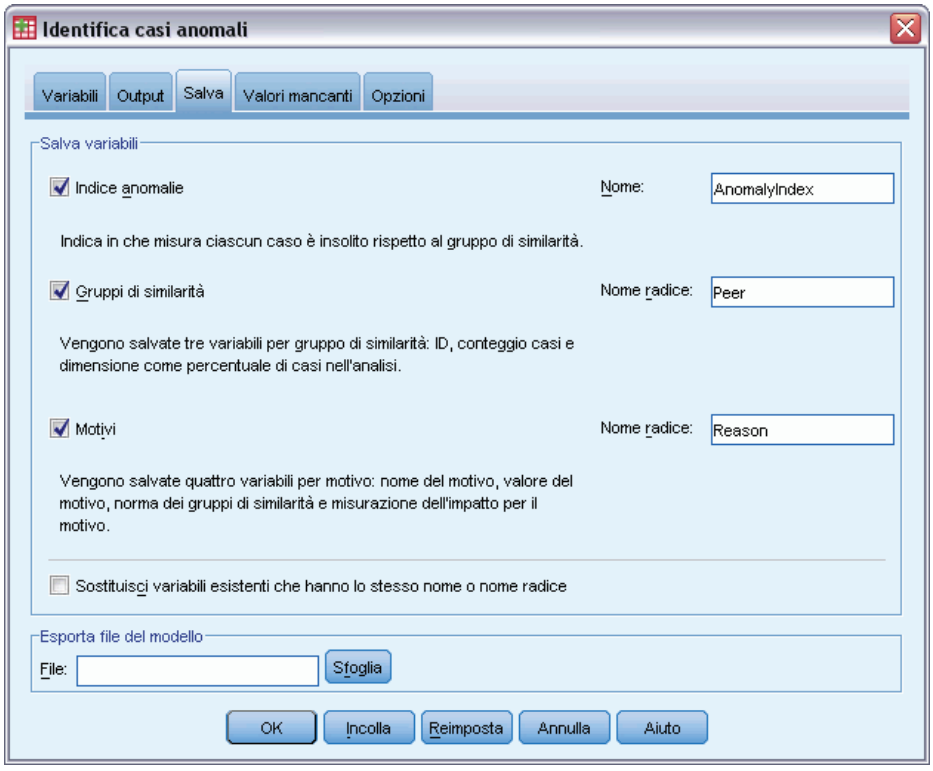
- **Norme dei gruppi equivalenti.** Questa opzione visualizza la tabella delle norme delle variabili continue se l'analisi utilizza questo tipo di variabili e la tabella delle norme delle variabili categoriche se l'analisi utilizza questo tipo di variabili. La prima tabella visualizza la media e la deviazione standard di ciascuna variabile continua in ciascun gruppo equivalente. La seconda tabella visualizza la modalità (ovvero la categoria più popolare) unitamente alla frequenza e alla frequenza percentuale di ciascuna variabile categoriale di ciascun gruppo

equivalente. La media della variabile continua e la modalità della variabile categorica vengono usate come norma nell'analisi.

- **Indici di anomalia.** Il riassunto Indice delle anomalie visualizza le statistiche descrittive per l'indice delle anomalie dei casi identificati come particolarmente anomali.
- **Occorrenza motivi per variabile di analisi.** La tabella visualizza per ciascun motivo la frequenza e la frequenza percentuale dell'occorrenza di ciascuna variabile come motivo. La tabella fornisce anche la statistica descrittiva dell'impatto di ciascuna variabile. Se il numero massimo di motivi è impostato su 0 nella scheda Opzioni, l'opzione non è disponibile.
- **Casi elaborati.** Il riepilogo di elaborazione dei casi visualizza i conteggi e i conteggi percentuali per tutti i casi dell'insieme di dati attivo, i casi inclusi ed esclusi dall'analisi, e i casi di ciascun gruppo equivalente.

Scheda Salva di Identifica casi anomali

Figura 5-4
Scheda Salva della finestra di dialogo Identifica casi anomali



Salva variabili. I controlli di questo gruppo permettono di salvare le variabili modello nell'insieme di dati attivo. È possibile anche scegliere di sostituire le variabili esistenti il cui nome è in conflitto con le variabili da salvare.

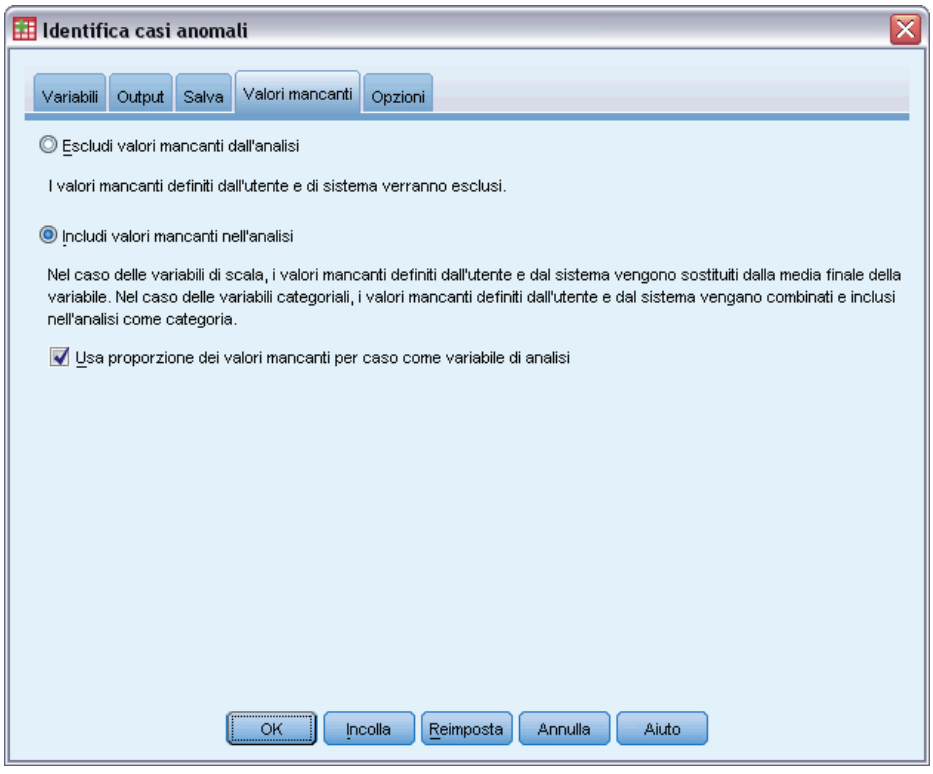
- **Indice di anomalia.** Salva il valore dell'indice di anomalia per ciascun caso nella variabile con un nome specifico.

- **Gruppi equivalenti.** Salva l'ID del gruppo equivalente, il conteggio dei casi e la dimensione percentuale di ogni caso nelle variabili con il nome radice specificato. Ad esempio se è stato specificato il nome radice *Peer*, vengono generate le variabili *Peerid*, *PeerSize* e *PeerPctSize*. *Peerid* è l'ID del gruppo equivalente del caso, *PeerSize* è la dimensione del gruppo e *PeerPctSize* è la dimensione percentuale del gruppo.
- **Motivi.** Salva gli insiemi di variabili dei motivi con il nome radice specificato. Gli insiemi di variabili di questo tipo comprendono il nome della variabile sotto forma di motivo, la misura dell'impatto della variabile, il valore della variabile e il valore della norma. Il numero di insiemi dipende dal numero di motivi specificato nella scheda Opzioni. Ad esempio se è stato specificato il nome radice *Reason*, vengono generate le variabili *ReasonVar_k*, *ReasonMeasure_k*, *ReasonValue_k* e *ReasonNorm_k*, dove *k* è il *k*° motivo. Questa opzione non è disponibile se il numero di motivi è impostato su 0.

Esporta file del modello. Permette di salvare il modello in formato XML.

Scheda Valori mancanti in Identifica casi anomali

Figura 5-5
Scheda Valori mancanti della finestra di dialogo Identifica casi anomali

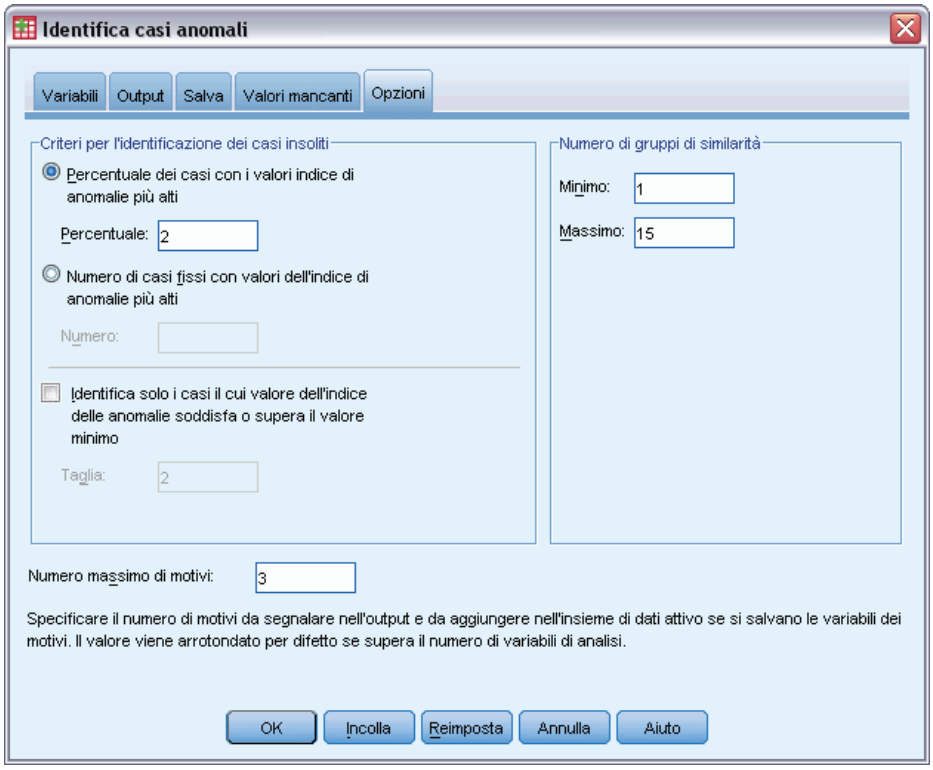


La scheda Valori mancanti permette di controllare la gestione dei valori mancanti definiti dall'utente e di sistema.

- **Escludi valori mancanti dall'analisi.** Esclude i casi con valori mancanti dall'analisi.
- **Includi valori mancanti nell'analisi.** I valori mancanti delle variabili continue vengono sostituiti con le medie totali corrispondenti, mentre le categorie mancanti delle variabili categoriche vengono raggruppate e considerate equivalenti a categorie valide. Le variabili elaborate vengono quindi utilizzate per l'analisi. È possibile anche richiedere la creazione di una variabile aggiuntiva che rappresenta la proporzione di variabili mancanti di ciascun caso e usarla per l'analisi.

Scheda Opzioni di Identifica casi anomali

Figura 5-6
Scheda Opzioni della finestra di dialogo Identifica casi anomali



Criteri per l'identificazione dei casi anomali. Queste opzioni permettono di specificare quanti casi includere nell'elenco delle anomalie.

- **Percentuale di casi con i valori indice di anomalia più alti.** Specificare un numero positivo uguale o minore di 100.
- **Numero fisso di casi con i valori indice di anomalia più alti.** Specificare un intero positivo uguale o minore del numero totale di casi nell'insieme di dati attivo usati per l'analisi.
- **Identifica solo i casi con valore indice di anomalia uguale o maggiore del valore minimo.** Specificare un intero non negativo. Un caso viene considerato anomalo se il suo indice di anomalia è uguale o maggiore del punto di riferimento specificato. Questa opzione viene generalmente usata insieme alle opzioni Percentuale di casi e Numero fisso di casi. Se si specifica, ad esempio, un numero fisso di 50 casi e un valore di riferimento pari a 2, l'elenco

delle anomalie comprenderà un massimo di 50 casi, ciascuno con un indice di anomalia uguale o maggiore di 2.

Numero di gruppi equivalenti. Questa procedura ricerca il numero migliore di gruppi equivalenti compreso tra i valori minimo e massimo. I valori devono essere interi positivi e il minimo deve sempre essere inferiore al massimo. Se i valori specificati sono uguali, la procedura utilizza un numero fisso di gruppi equivalenti.

Nota: a seconda della variabilità dei dati, può accadere anche che il numero di gruppi equivalenti supportato dai dati sia inferiore al numero minimo specificato. In questo caso è possibile che la procedura generi un numero minore di gruppi equivalenti.

Numero massimo di motivi. Un motivo è costituito dalla misura di impatto della variabile, dal nome della variabile corrispondente al motivo, dal valore della variabile e dal valore del gruppo equivalente corrispondente. Specificare un numero intero non negativo. Se questo valore è uguale o maggiore del numero di variabili elaborate usate nell'analisi, vengono visualizzate tutte le variabili.

Funzioni aggiuntive del comando DETECTANOMALY

Il linguaggio della sintassi dei comandi consente inoltre di:

- Omettere alcune variabili dell'insieme di dati dall'analisi senza specificare esplicitamente tutte le variabili dell'analisi (con il sottocomando `EXCEPT`).
- Specificare una rettifica per bilanciare l'influenza delle variabili continue e categoriali (con la parola chiave `MLWEIGHT` nel sottocomando `CRITERIA`).

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

Capitolo
6

Categorizzazione ottimale

La procedura Categorizzazione ottimale discretizza una o piu variabili di scala (d'ora in avanti dette **variabili di input per la categorizzazione**) distribuendo i valori di ogni variabile in intervalli. La formazione di intervalli risulta migliore rispetto a una variabile guida categoriale che esegue la "supervisione" del processo di categorizzazione, in quanto sarà possibile utilizzare gli intervalli anziché i valori dei dati originali per ulteriori analisi.

Esempi. La riduzione del numero di valori distinti assunti da una variabile puo essere utile a diversi fini, fra cui:

- Requisiti di dati per altre procedure. E possibile trattare le variabili discretizzate come categoriali per utizzarle in procedure che richiedono variabili di questo tipo. Ad esempio, la procedura Tavole di contingenza richiede che tutte le variabili siano categoriali.
- Riservatezza dei dati. L'inserimento in report di valori con categorizzazione anziché dei valori effettivi può essere utile per salvaguardare la riservatezza delle sorgenti dati. La procedura Categorizzazione ottimale è in grado di pilotare la scelta degli intervalli.
- Velocita. Alcune procedure vengono svolte con maggiore efficienza quando si lavora con un numero ridotto di valori distinti. Ad esempio, la velocita di Regressione logistica multinominale aumenta se si utiizzano variabili discretizzate.
- Identificazione della separazione dei dati completa o quasi completa.

Categorizzazione ottimale e Categorizzazione visuale. La finestra di dialogo Categorizzazione visuale offre vari metodi automatici di creare intervalli senza utilizzare una variabile guida. Le regole per lo svolgimento del processo senza supervisione sono utili per generare statistiche descrittive, quali tabelle di frequenza, mentre la categorizzazione ottimale offre risultati migliori quando lo scopo finale e creare un modello predittivo.

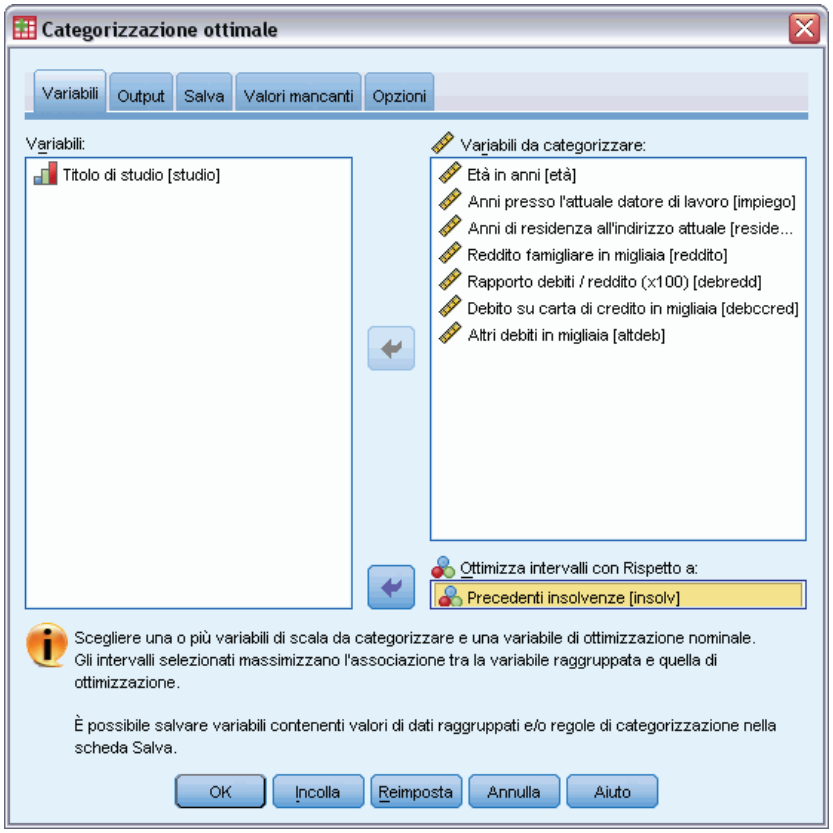
Output. La procedura produce come risultato tabelle di punti di divisione per gli intervalli e statistiche descrittive per ogni variabile di input per la categorizzazione. È inoltre possibile salvare nell'insieme di dati attivo nuove variabili contenenti i valori categorizzati delle variabili di input per la categorizzazione e salvare le regole di categorizzazione come sintassi di comando per la discretizzazione di nuovi dati.

Dati. Per questa procedura e previsto che le variabili di input per la categorizzazione siano variabili di scala numeriche. La variabile guida deve essere categoriale e puo essere sia una stringa che numerica.

Per eseguire la categorizzazione ottimale

Dai menu, scegliere:
Trasforma > Categorizzazione ottimale...

Figura 6-1
Scheda Variabili della finestra di dialogo Categorizzazione ottimale

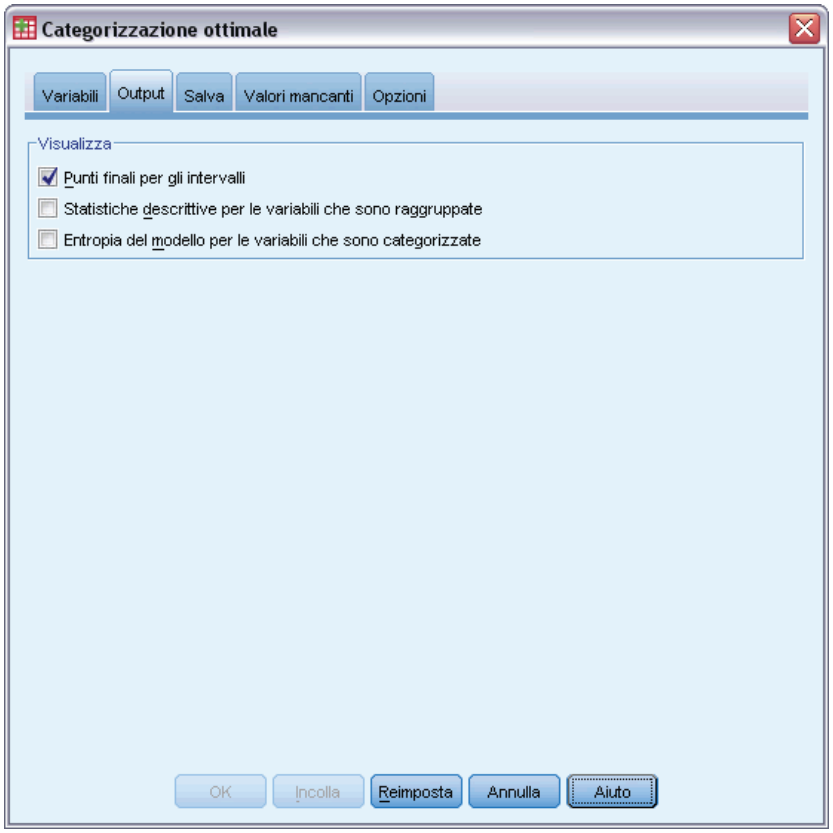


- Selezionare una o piu variabili di input per la categorizzazione.
- Selezionare una variabile guida.

Le variabili contenenti i valori dei dati con binning non vengono generate per default. Utilizzare la scheda Salva per salvare tali variabili.

Output della categorizzazione ottimale

Figura 6-2
Scheda Output della finestra di dialogo Categorizzazione ottimale

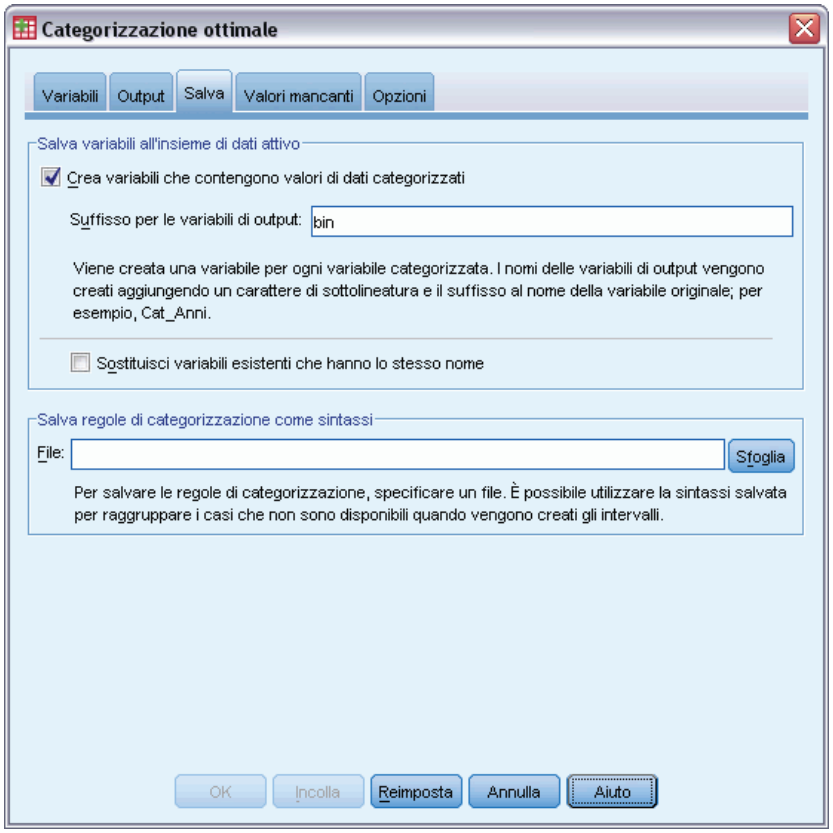


La scheda Output controlla la visualizzazione dei risultati.

- **Punti finali per gli intervalli.** Visualizza l'insieme dei punti finali per ogni variabile di input per la categorizzazione.
- **Statistiche descrittive per le variabili categorizzate.** Per ogni variabile di input per la categorizzazione, questa opzione visualizza il numero di casi con valori validi, il numero di casi con valori mancanti, il numero di valori validi distinti e il valore minimo e massimo. Per la variabile guida, questa opzione visualizza la distribuzione di classe per ogni variabile di input correlata per la categorizzazione.
- **Modello di entropia per le variabili categorizzate.** Per ogni variabile di input per la categorizzazione, questa opzione visualizza una misura della precisione predittiva della variabile rispetto alla variabile guida.

Salva di Categorizzazione ottimale

Figura 6-3
Scheda Salva della finestra di dialogo Categorizzazione ottimale

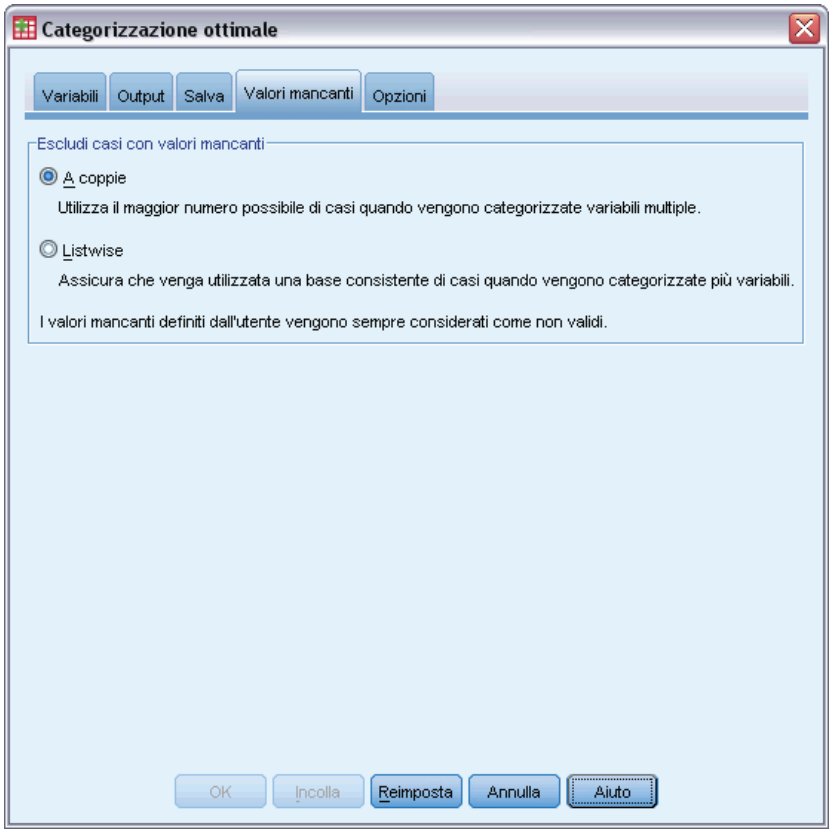


Salva le variabili in un insieme di dati attivo. Per ulteriori analisi e possibile utilizzare le variabili contenenti i valori dei dati con categorizzazione anziché le variabili originali.

Salva regole di categorizzazione come sintassi. Genera la sintassi dei comandi utilizzabile per eseguire la categorizzazione di altri insiemi di dati. Le regole di ricodifica si basano sui punti di divisione determinati dall'algoritmo di categorizzazione.

Valori mancanti di Categorizzazione ottimale

Figura 6-4
Scheda Valori mancanti della finestra di dialogo Categorizzazione ottimale

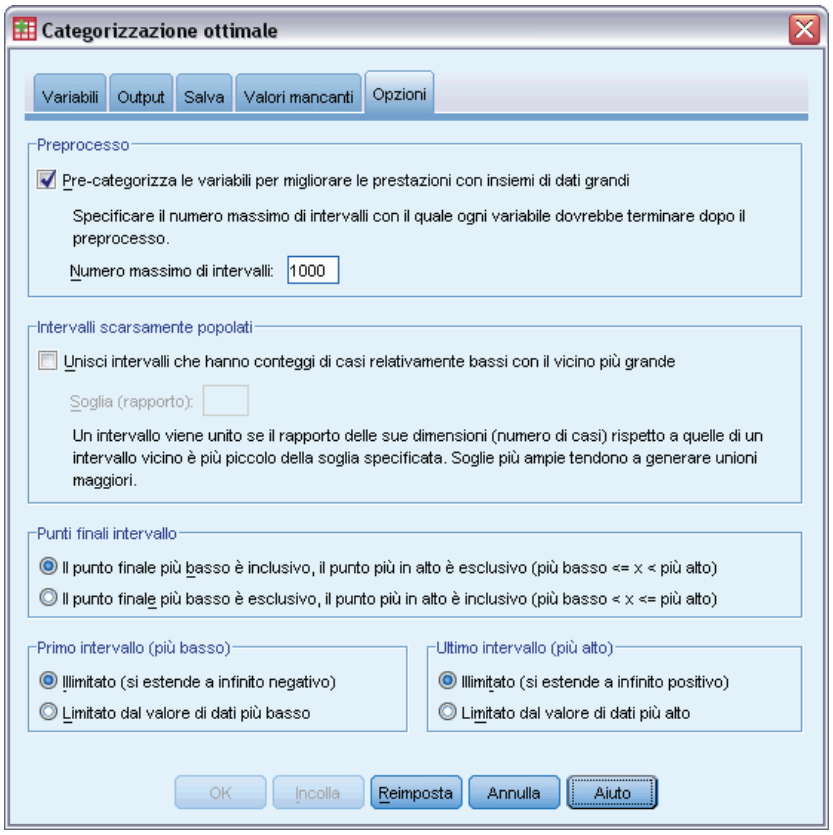


La scheda Valori mancanti specifica se i valori mancanti vengono gestiti utilizzando l'eliminazione listwise o pairwise. I valori mancanti definiti dall'utente vengono sempre considerati come non validi. Quando si ricodificano i valori della variabile originale nella nuova variabile, i valori mancanti definiti dall'utente vengono convertiti in valori mancanti definiti dal sistema.

- **Paiwise (Analisi dei dati mancanti).** Questa opzione agisce su tutte le coppie variabile guida/variabile di input per la categorizzazione. La procedura utilizza tutti i casi con valori non mancanti presenti nella variabile guida e nella variabile di input per la categorizzazione.
- **Listwise.** Questa funzione agisce su tutte le variabili specificate nella scheda Variabili. Se ad un caso manca una variabile, viene escluso l'intero caso.

Opzioni di Categorizzazione ottimale

Figura 6-5
Scheda Opzioni della finestra di dialogo Categorizzazione ottimale



Preprocesso. Utilizzando variabili di input per la categorizzazione con molti valori distinti nel processo di "pre-categorizzazione" e possibile ridurre i tempi di elaborazione senza pregiudicare la qualità degli intervalli finali. Il numero massimo di intervalli fornisce il limite superiore del numero di intervalli creati. Quindi, se si specifica 1000 come massimo, ma una variabile di input per la categorizzazione ha meno di 1000 valori distinti, il numero di intervalli per i quali viene eseguito il preprocesso creati per la variabile di input per la categorizzazione sarà uguale al numero di valori distinti presenti nella variabile di input per la categorizzazione.

Intervalli scarsamente popolati. In qualche caso la procedura puo dare luogo a intervalli con pochi casi. Con la strategia seguente si eliminano i pseudo punti di divisione:

- Per una data variabile, si supponga che l'algoritmo trovi n punti di divisione $_{\text{finali}}$ e quindi n intervalli $_{\text{finali}}$+1. Per gli intervalli $i = 2, ..., n_{\text{finali}}$ (dall'intervallo con il secondo valore più basso all'intervallo con il secondo valore più alto), calcolare

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

dove *sizeof(b)* corrisponde al numero di casi presenti nell'intervallo.

- Quando questo valore è inferiore alla soglia di unione specificata, b_i viene considerato scarsamente popolato e viene unito con b_{i-1} o b_{i+1}, a seconda di quale presenta la minore entropia delle informazioni di classe.

La procedura prevede un singolo passaggio attraverso gli intervalli.

Punti finali intervallo. Questa opzione specifica le modalità di definizione del limite inferiore di un intervallo. Dato che il valore dei punti di divisione viene determinato automaticamente, si tratta essenzialmente di una questione di preferenza.

Primo intervallo (più basso) / Ultimo intervallo (più alto). Queste opzioni specificano le modalità di definizione dei punti di divisione minimo e massimo per ogni variabile di input per la categorizzazione. In generale, nella procedura si presume che le variabili di input per la categorizzazione possano assumere qualsiasi valore fra i numeri reali, ma se sussistono ragioni teoriche o pratiche per limitare l'intervallo, specificare come limiti il valore più basso e il più alto.

Opzioni aggiuntive del comando OPTIMAL BINNING

Il linguaggio della sintassi dei comandi consente inoltre di:

- Eseguire la categorizzazione senza supervisione tramite il metodo delle frequenze uguali (utilizzando il sottocomando `CRITERIA`).

Per informazioni dettagliate sulla sintassi, vedere *Command Syntax Reference*.

Parte II: Esempi

Capitolo
7

Convalida dati

La procedura Convalida dati permette di identificare casi, variabili e valori dati sospetti e non validi.

Convalida di un database medico

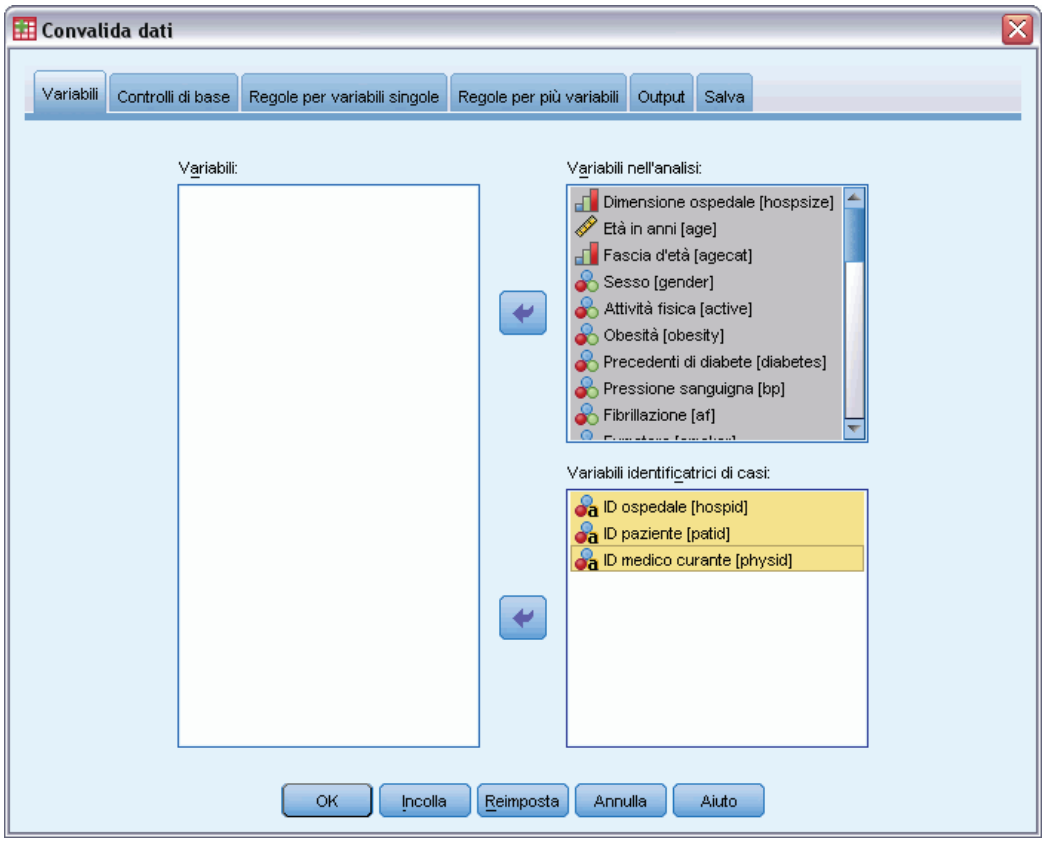
L'analisi assunto da un istituto medico ha l'esigenza di garantire la qualità delle informazioni contenute nel sistema. Questo processo implica controllare i valori e le variabili nonché preparare report per il responsabile del team di inserimento dati.

L'ultima versione del database è contenuta in *stroke_invalid.sav*. Per ulteriori informazioni, vedere l'argomento File di esempio in l'appendice A a pag. 137. Usare la procedura Convalida dati per acquisire le informazioni necessarie per creare il report. La sintassi per la creazione di queste analisi è contenuta in *validatedata_stroke.sps*.

Esecuzione dei controlli di base

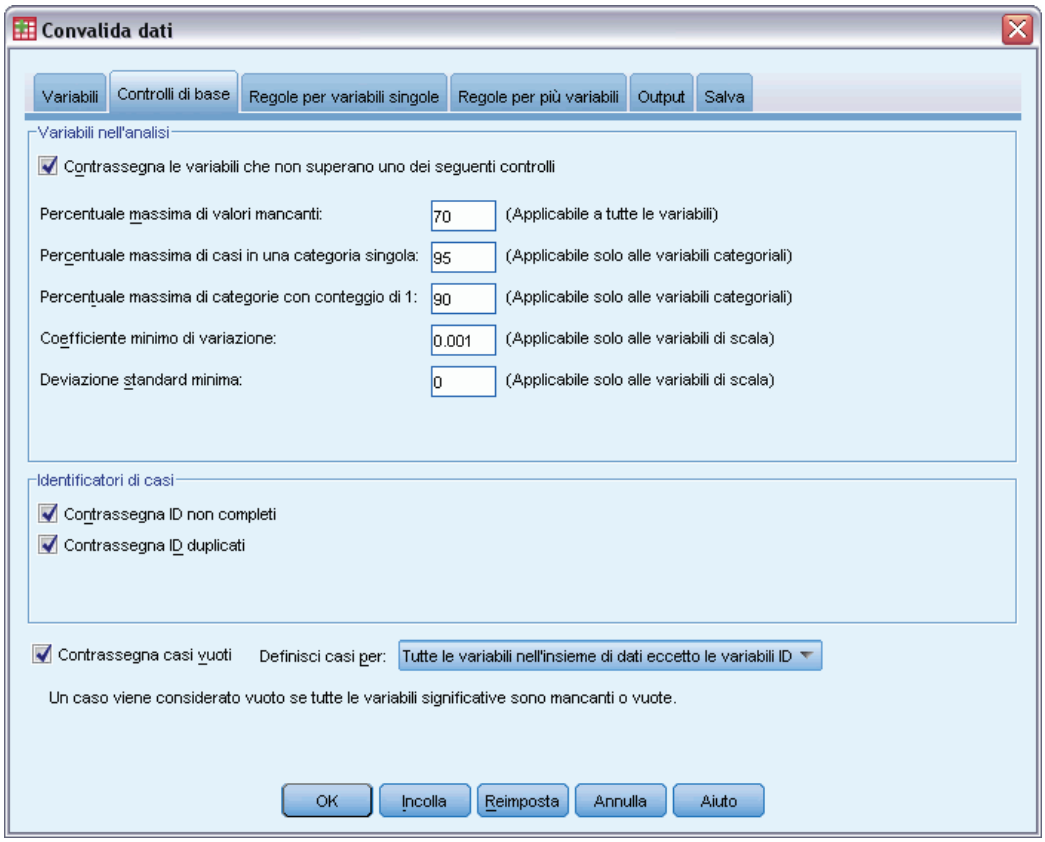
- Per eseguire un'analisi di convalida dei dati, dai menu scegliere:
  Dati > Convalida > Convalida dati...

Figura 7-1
Scheda Variabili della finestra di dialogo Convalida dati



- Selezionare le variabili da *Dimensioni ospedale* e *Età in anni* a *Indice di Barthel registrato dopo 6 mesi* come variabili dell'analisi.
- Selezionare *ID ospedale*, *ID paziente* e *ID medico responsabile* come variabile di identificazione dei casi.
- Fare clic sulla scheda Controlli di base.

Figura 7-2
Scheda Controlli di base della finestra di dialogo Convalida dati



Le impostazioni predefinite sono quelle da utilizzare per l'operazione.

► Fare clic su OK.

Avvisi

Figura 7-3
Avvisi

Avvisi

Alcuni o tutti i risultati richiesti non verranno visualizzati perché tutti i casi, le variabili o i valori dei dati hanno soddisfatto i controlli richiesti.

Le variabili dell'analisi soddisfano i controlli di base e non ci sono casi vuoti. Quindi, viene visualizzato un avviso che spiega che i controlli non hanno prodotto alcun output.

Identificatori incompleti

Figura 7-4
Identificatori di casi incompleti

Identificatori incompleti

	Identificatore		
Caso	ID ospedale	ID paziente	ID medico curante
288	OZN		125304
573		6137798782	790697
774		2322241867	176466

Se le variabili di identificazione dei casi contengono valori mancanti, non è possibile identificare correttamente il caso. Nel file di dati dell'esempio mancano *ID paziente* nel caso 288 e *ID ospedale* nei casi 573 e 774.

Identificatori duplicati

Figura 7-5
Identificatori di casi duplicati (vengono visualizzati solo i primi 11)

Identificatori duplicati

Gruppo degli identificatori duplicati	Numero di duplicati	Casi con identificatori duplicati	Identificatore		
			ID ospedale	ID paziente	ID medico curante
1	2	10, 11	PBW	1406462419	355184
2	2	14, 15	PBW	2191527525	355184
3	2	21, 22	PBW	7237535360	616528
4	2	28, 29	NHV	4592215163	942982
5	2	30, 31	NHV	7628592330	371884
6	2	64, 65	NHV	0300750006	371884
7	2	83, 84	QWS	4590625286	215041
8	2	86, 87	QWS	6272818258	817329
9	2	96, 97	QWS	1959349605	215041
10	3	100, 101, 102	QWS	5856145337	817329
11	3	104, 105, 106	QWS	1543897849	817329

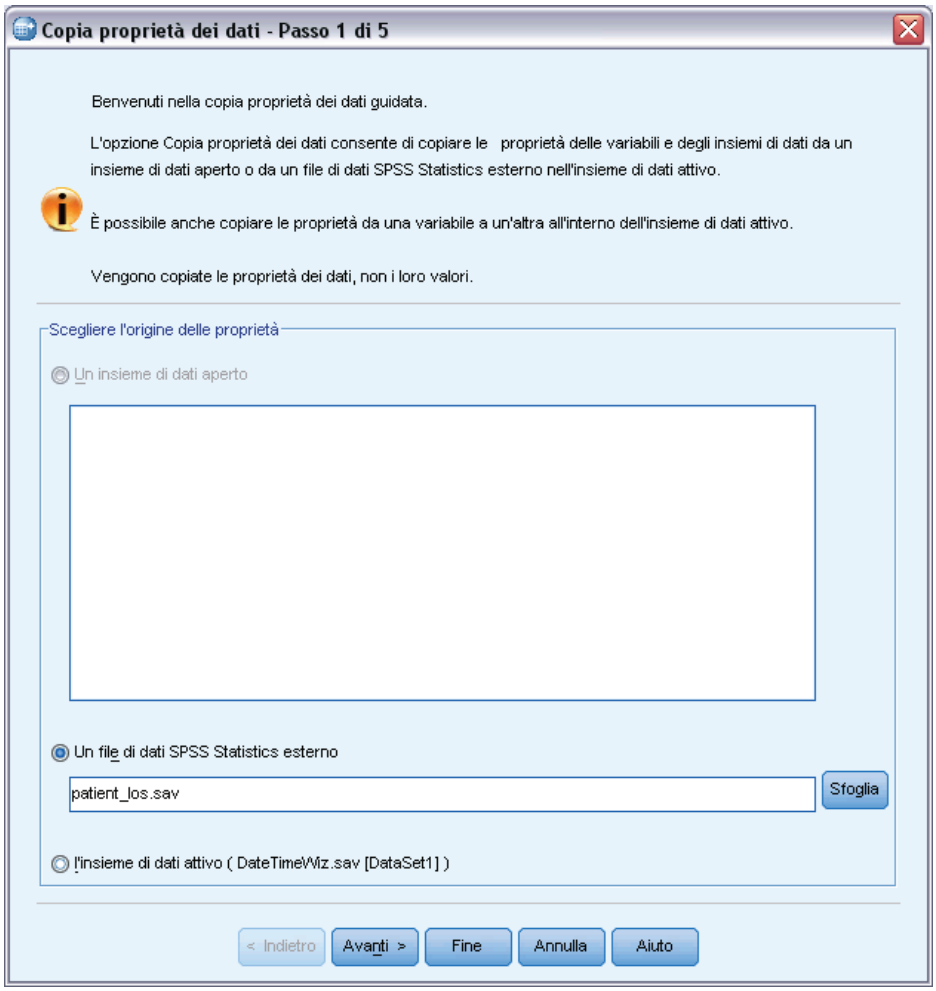
I casi devono essere identificati in modo univoco dalla combinazione di valori delle variabili di identificazione. Nell'esempio vengono visualizzate le prime 11 voci della tabella degli identificatori duplicati. Questi duplicati sono riferiti a pazienti con più eventi che sono stati immessi come casi diversi per ciascun evento. Poiché è possibile raggruppare queste informazioni su un'unica riga, è necessario pulire questi casi.

Copia e uso delle regole di un altro file

L'analista nota che le variabili del file di dati sono simili a quelle di un altro progetto. Le regole di convalida definite per il progetto vengono salvate come proprietà del file di dati correlato e possono essere applicate al file di dati semplicemente copiando le proprietà dati del file.

▶ Per copiare le regole da un altro file, selezionare dai menu:
Dati > Copia proprietà dei dati...

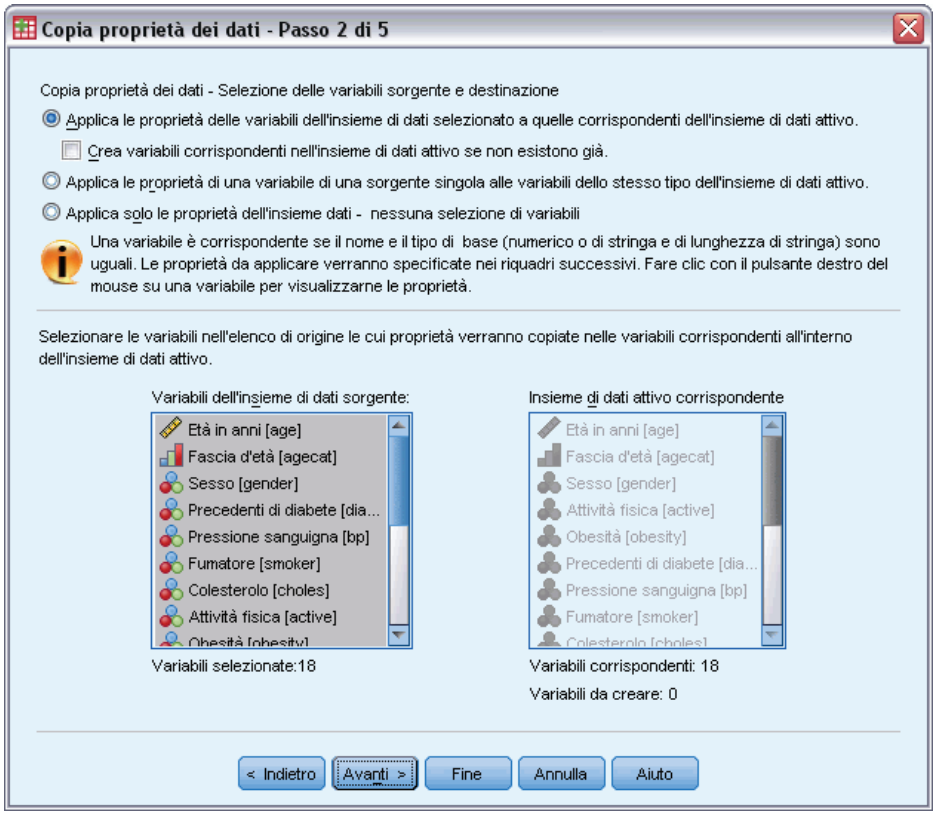
Figura 7-6
Copia proprietà dei dati, Passo 1 (finestra di benvenuto)



▶ Scegliere di copiare le proprietà da un file di dati IBM® SPSS® Statisticsesterno, *patient_los.sav*. Per ulteriori informazioni, vedere l'argomento File di esempio in l'appendice A a pag. 137.

▶ Fare clic su Avanti.

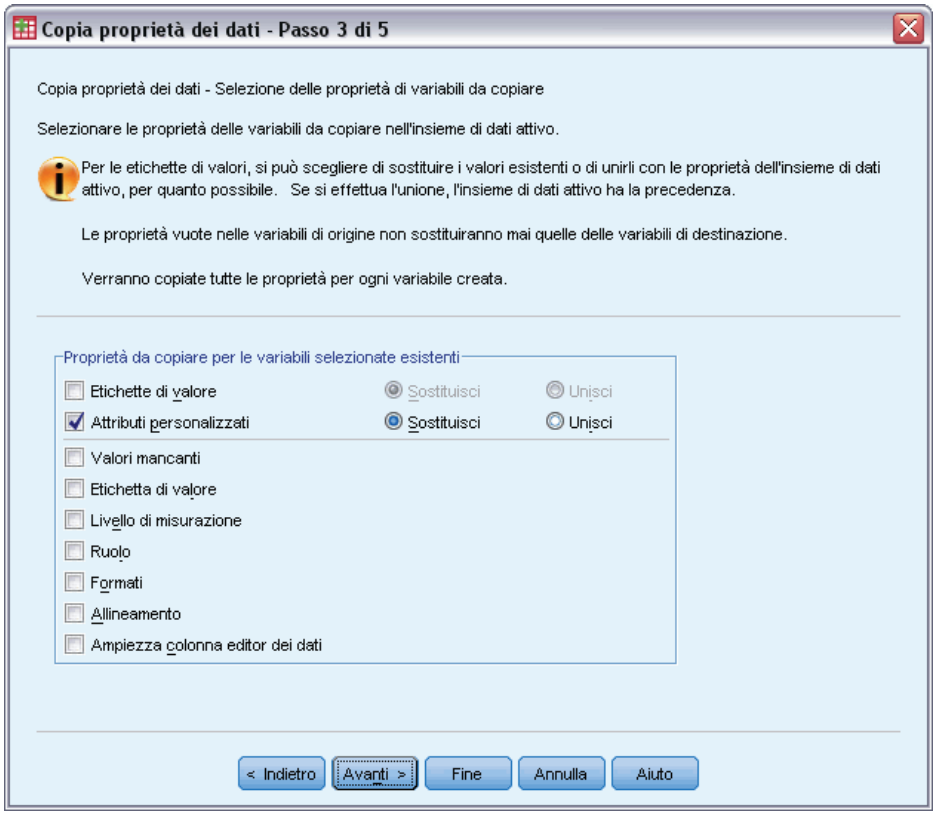
Figura 7-7
Copia proprietà dei dati, Passo 2 (selezione delle variabili)



Queste sono le variabili le cui proprietà devono essere copiate da *patient_los.sav* nelle variabili corrispondenti di *stroke_invalid.sav*.

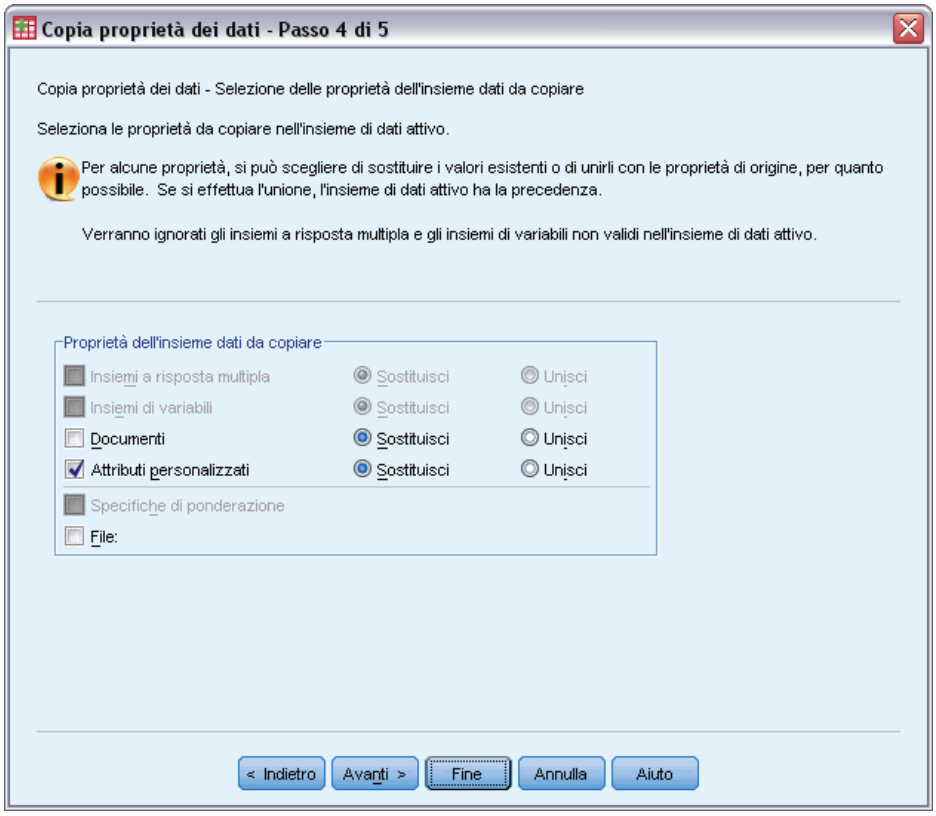
▶ Fare clic su Avanti.

Figura 7-8
Copia proprietà dei dati, Passo 3 (selezione delle proprietà delle variabili)



- Deselezionare tutte le proprietà ad eccezione di Attributi personalizzati.
- Fare clic su Avanti.

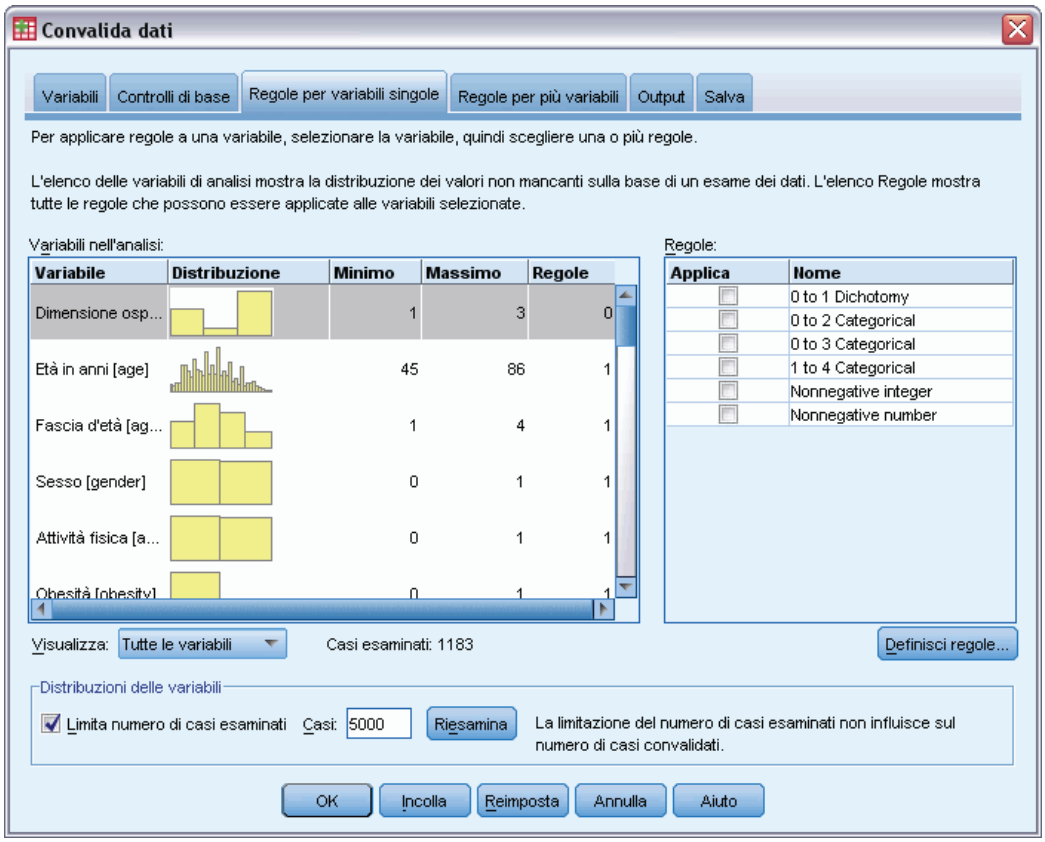
Figura 7-9
Copia proprietà dei dati, Passo 4 (selezione delle proprietà dei file di dati)



- Selezionare Attributi personalizzati.
- Fare clic su Fine.

A questo punto è possibile riutilizzare le regole di convalida.

Figura 7-10
Scheda Regole per variabili singole della finestra di dialogo Convalida dati

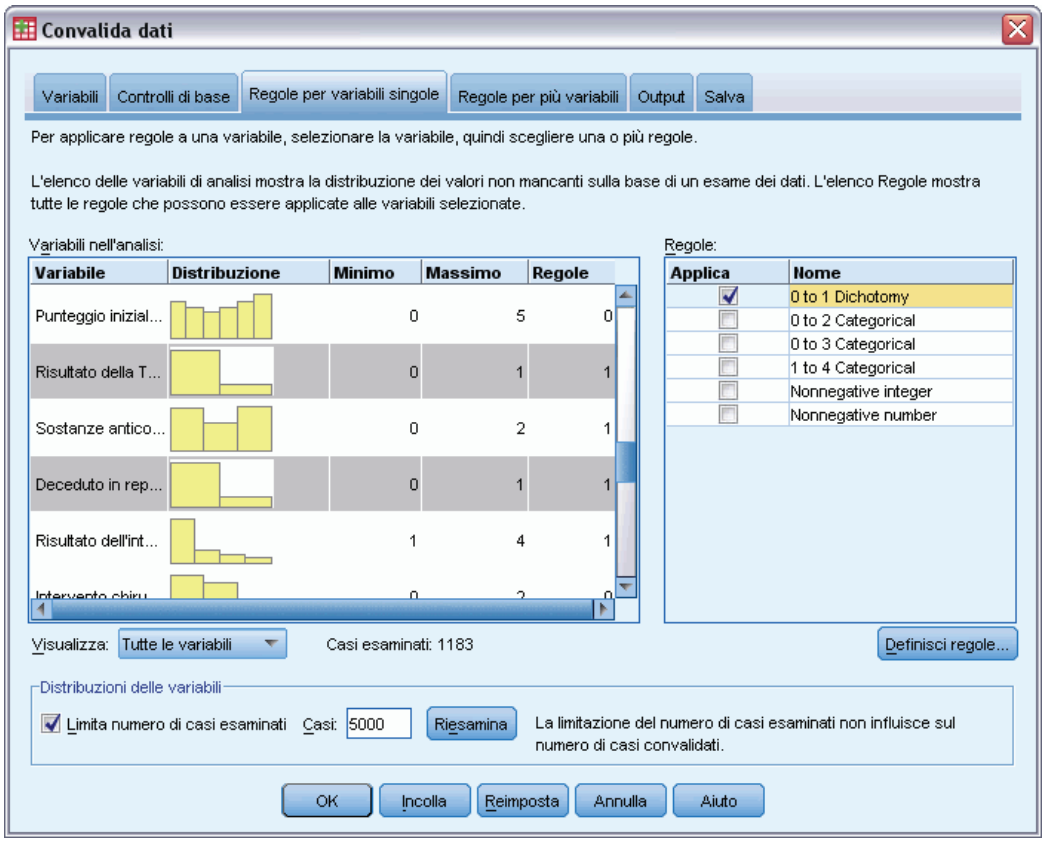


- Per convalidare i dati di *stroke_invalid.sav* utilizzando le regole copiate, fare clic sul pulsante Richiama finestra nella barra degli strumenti e selezionare Convalida dati.
- Fare clic sulla scheda Regole per variabili singole.

L'elenco Variabili dell'analisi visualizza le variabili selezionate nella scheda Variabili, alcune informazioni riassuntive sulle loro distribuzioni e il numero di regole associate a ciascuna variabile. Le variabili le cui proprietà sono state copiate da *patient_los.sav* sono associate a regole.

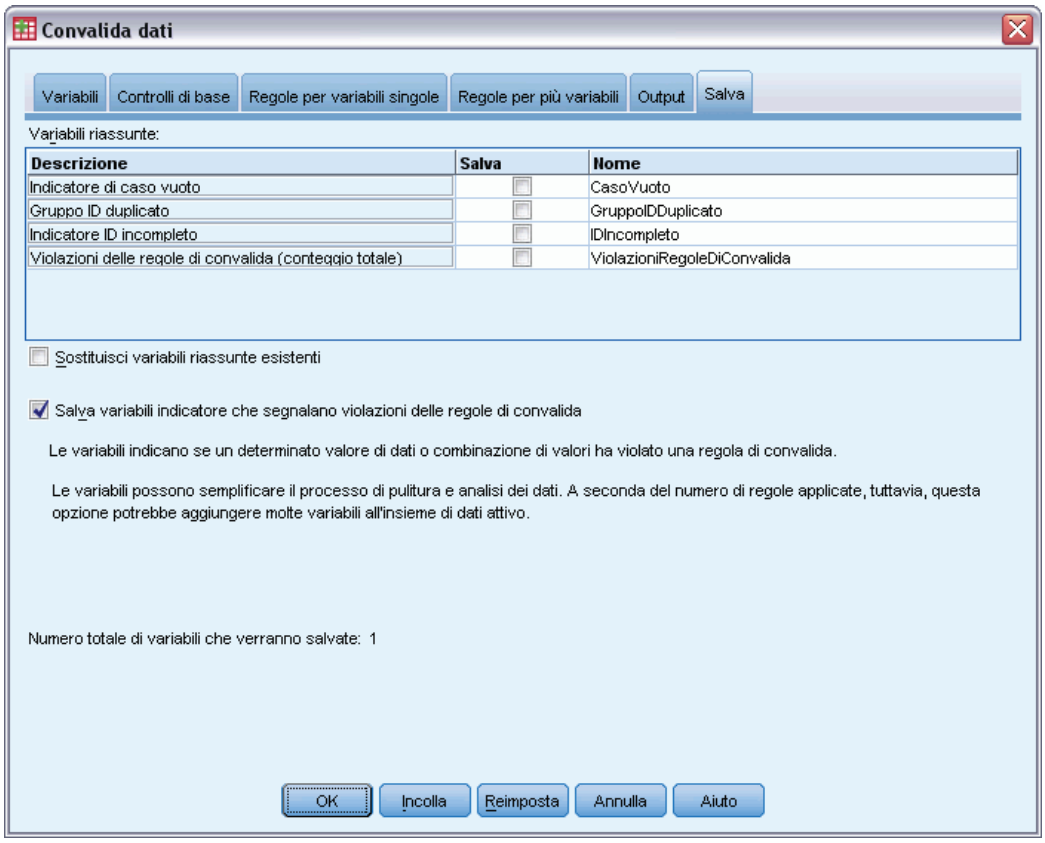
L'elenco Regole visualizza le regole per variabili singole disponibili nel file di dati. Queste regole sono state tutte copiate da *patient_los.sav*. Tuttavia, è utile notare che alcune delle regole si riferiscono a variabili che non hanno corrispondenze nell'altro file di dati.

Figura 7-11
Scheda Regole per variabili singole della finestra di dialogo Convalida dati



- Selezionare *Fibrillazione atriale*, *Attacchi ischemici temporanei precedenti*, *Risultato scansione CAT* e *Deceduto in ospedale*, quindi applicare la regola Dicotomia da 0 a 1.
- Applicare Categoriale da 0 a 3 a *Riabilitazione post-evento*.
- Applicare Categoriale da 0 a 2 a *Chirurgia preventiva post-evento*.
- Applicare Valore interno non negativo a *Durata della riabilitazione*.
- Applicare Categoriale da 1 a 4 a tutti i campi compresi tra *Indice di Barthel registrato dopo 1 mese* e *Indice di Barthel registrato dopo 6 mesi*.
- Fare clic sulla scheda Salva.

Figura 7-12
Scheda Salva della finestra di dialogo Convalida dati



- Selezionare Salva variabili indicatore che registrano tutte le violazioni alle regole di convalida. Questa operazione permette di collegare più facilmente il caso alla variabile che provoca al violazione alla regola per variabili singole.
- Fare clic su OK.

Descrizioni delle regole

Figura 7-13
Descrizioni delle regole

Regola	Descrizione
Nonnegative integer	Tipo: Numerico Dominio: Intervallo Contrassegna valori mancanti definiti dall'utente: No Contrassegna valori mancanti definiti dal sistema: Sì Minimo: 0 Contrassegna valori compresi nell'intervallo: No Contrassegna valori non interi compresi nell'intervallo: Sì $VD.SRule[5]: Regola
0 to 1 Dichotomy	Tipo: Numerico Dominio: Elenco Contrassegna valori mancanti definiti dall'utente: No Contrassegna valori mancanti definiti dal sistema: Sì Elenco: 0, 1 $VD.SRule[1]: Regola
1 to 4 Categorical	Tipo: Numerico Dominio: Elenco Contrassegna valori mancanti definiti dall'utente: No Contrassegna valori mancanti definiti dal sistema: Sì Elenco: 1, 2, 3, 4 $VD.SRule[4]: Regola

Vengono visualizzate le regole violate

La tabella Descrizioni delle regole visualizza la descrizione delle regole violate. Questa funzione è molto utile per controllare molte regole di convalida.

Riepilogo variabili

Figura 7-14
Riepilogo variabili

	Regola	Numero di violazioni
agecat	1 to 4 Categorical	1
	Totale	1
gender	0 to 1 Dichotomy	1
	Totale	1
angina	0 to 1 Dichotomy	1
	Totale	1
time	Nonnegative integer	2
	Totale	2
doa	0 to 1 Dichotomy	1
	Totale	2

La tabella Riassunto variabili elenca le variabili che hanno violato almeno una regola di convalida, le regole che sono state violate e il numero di violazioni per regole e variabile.

Report dei casi

Figura 7-15
Report dei casi

	Violazioni delle regole	Identificatore		
Caso	Variabile singola[a]	hospid	patid	physid
175	0 to 1 Dichotomy (1)	OZN	0333204686	883285
274	0 to 1 Dichotomy (1)	OZN	1038840465	103254
310	Nonnegative integer (1)	OZN	2090290204	883285
437	0 to 1 Dichotomy (1)	WPA	2349729006	723384
752	Nonnegative integer (1)	GFG	4993307441	828754
1173	1 to 4 Categorical (1)	ALK	8737661990	185787

a. Il numero di variabili che ha violato la regola è mostrato dopo ciascuna regola

La tabella Report dei casi elenca i casi (per numero e ID) che hanno violato almeno una regola di convalida, le regole che sono state violate e il numero di volte in cui la regola è stata violata dal caso. I valori non validi vengono visualizzati nell'Editor dei dati.

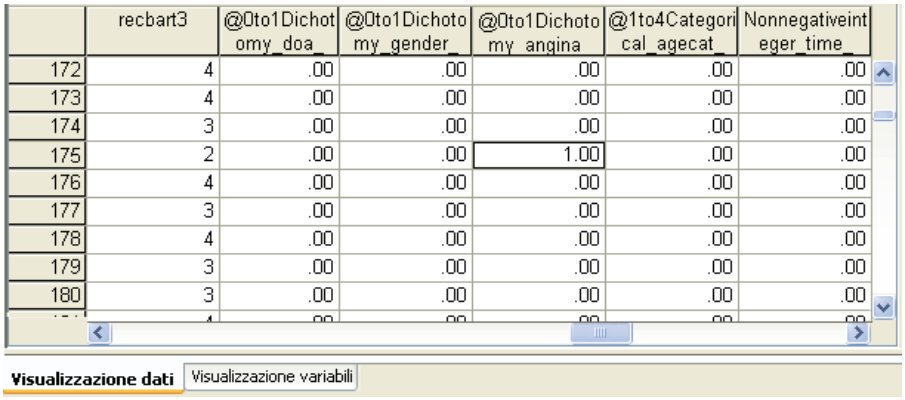
Figura 7-16
Editor dei dati con indicatori delle violazioni regole salvati

	recbart3	@0to3Categorical_clotsolv_	@0to3Categorical_rehab_	@0to1Dichotomy_obesity	@0to1Dichotomy_dhosp_	@0to1Dichotomy_tia_
1	4	.00	.00	.00	.00	
2	4	.00	.00	.00	.00	
3	1	.00	.00	.00	.00	
4	4	.00	.00	.00	.00	
5	3	.00	.00	.00	.00	
6	4	.00	.00	.00	.00	
7	4	.00	.00	.00	.00	
8	4	.00	.00	.00	.00	

Visualizzazione dati | Visualizzazione variabili

Viene creata una variabile indicatore per ciascuna applicazione della regola di convalida. Quindi, *@0to3Categorical_clotsolv_* rappresenta l'applicazione della regola di convalida per variabili singole Categoriale da 0 a 3 alla variabile *Trattati con anticoagulanti*. Il modo più semplice per stabilire quale valore della variabile di un caso non è valido, è quello di esaminare i valori degli indicatori. Un valore pari a 1 indica che il valore della variabile associato non è valido.

Figura 7-17
Editor dei dati con indicatore di violazione delle regole per il caso 175

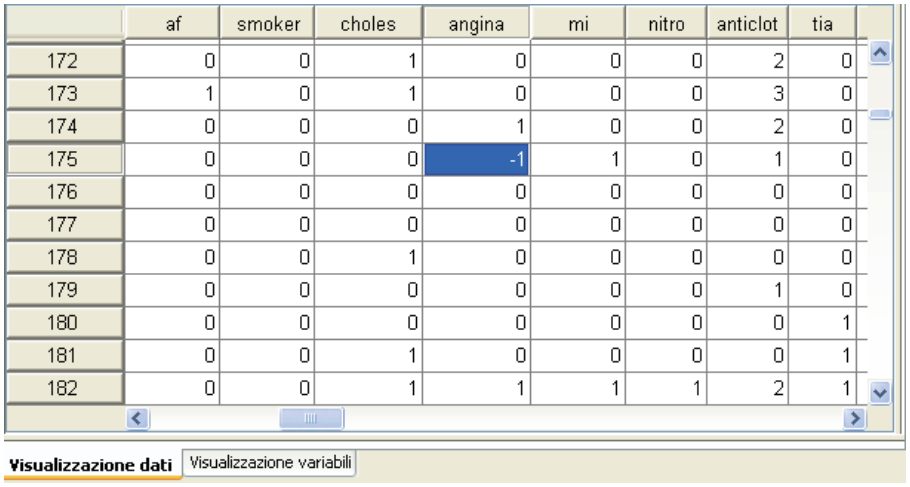


	recbart3	@0to1Dichotomy_doa_	@0to1Dichotomy_gender_	@0to1Dichotomy_angina	@1to4Categorical_agecat_	Nonnegativeinteger_time_
172	4	.00	.00	.00	.00	.00
173	4	.00	.00	.00	.00	.00
174	3	.00	.00	.00	.00	.00
175	2	.00	.00	1.00	.00	.00
176	4	.00	.00	.00	.00	.00
177	3	.00	.00	.00	.00	.00
178	4	.00	.00	.00	.00	.00
179	3	.00	.00	.00	.00	.00
180	3	.00	.00	.00	.00	.00

Visualizzazione dati | Visualizzazione variabili

Passare al caso 175, ovvero al primo caso con una violazione a una regola. Per rendere l'operazione più veloce, ricercare gli indicatori associati alle variabili nella tabella Riassunto variabili. Si nota che *Precedenti di angina* contiene un valore non valido.

Figura 7-18
Editor dei dati con valore non valido per Precedenti di angina



	af	smoker	choles	angina	mi	nitro	anticlot	tia
172	0	0	1	0	0	0	2	0
173	1	0	1	0	0	0	3	0
174	0	0	0	1	0	0	2	0
175	0	0	0	-1	1	0	1	0
176	0	0	0	0	0	0	0	0
177	0	0	0	0	0	0	0	0
178	0	0	1	0	0	0	0	0
179	0	0	0	0	0	0	1	0
180	0	0	0	0	0	0	0	1
181	0	0	1	0	0	0	0	1
182	0	0	1	1	1	1	2	1

Visualizzazione dati | Visualizzazione variabili

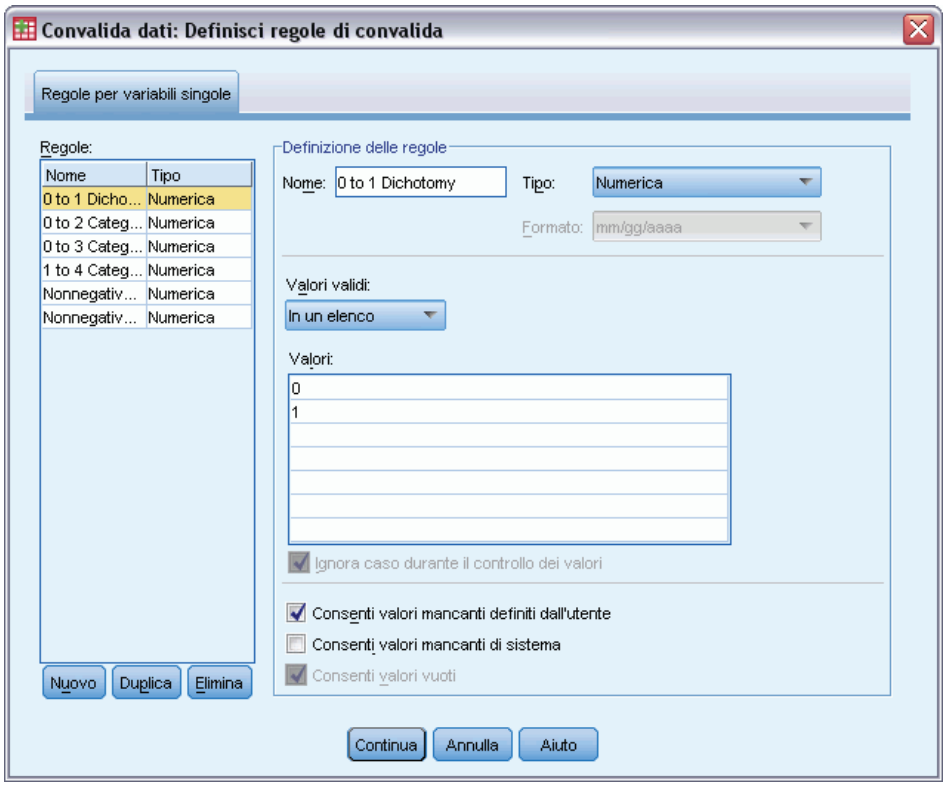
Precedenti di angina contiene il valore –1. Benché questo valore sia un valore mancante valido per le variabili Cura e Risultato nel file di dati, non è valido perché in questo caso per i valori dei precedenti del paziente non sono stati definiti valori mancanti definiti dall'utente.

Definizione di regole personalizzate

Le regole copiate da *patient_los.sav* sono state molto utili, ma per poter terminare il lavoro è necessario definire anche altre regole. Oltre a ciò è necessario tener presente che i pazienti che risultano deceduti all'arrivo vengono spesso classificati come deceduti in ospedale. Poiché le regole di convalida per variabili singole non riescono a rilevare questa situazione, è necessario definire regole per più variabili.

- Fare clic su Richiama finestra nella barra degli strumenti e selezionare Convalida dati.
- Fare clic sulla scheda Regole per variabili singole. Devono essere definite regole per *Dimensioni ospedale*, oltre alle variabili per la misurazione dei punteggi Ranking e le variabili che corrispondono agli indici di Barthel non registrati.
- Fare clic su Definisci regole.

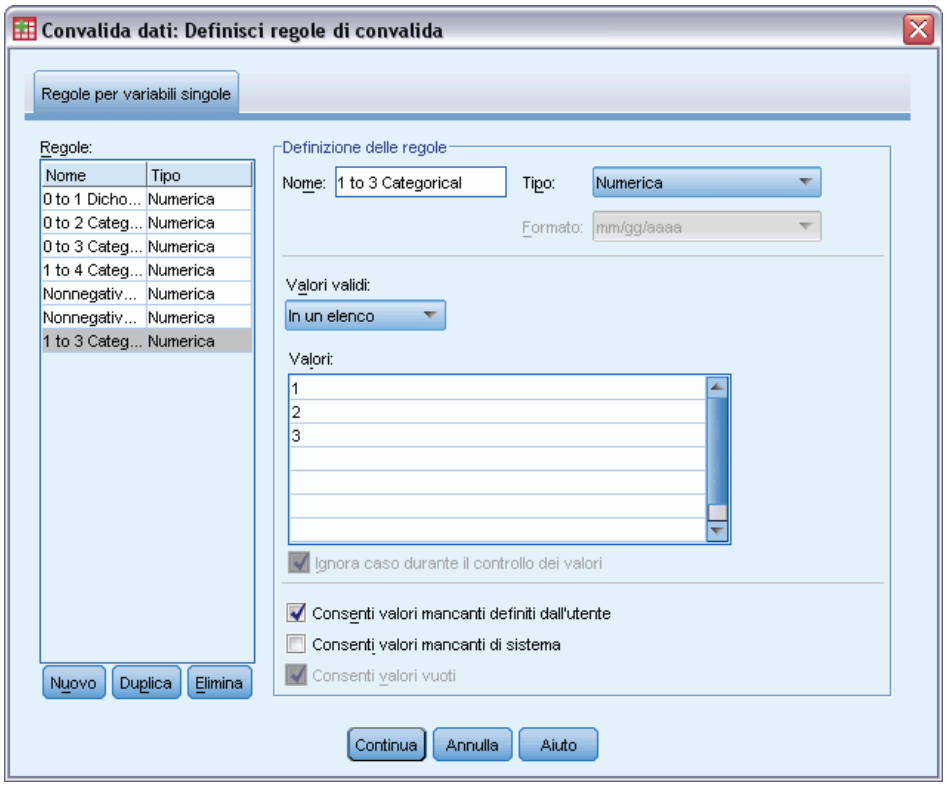
Figura 7-19
Scheda Regole per variabili singole della finestra di dialogo Definisci regole di convalida



Le regole correntemente definite vengono visualizzate con Dicotomia da 0 a 1 selezionata nell'elenco Regole e le proprietà delle regole visualizzate nel gruppo Definizione delle regole.

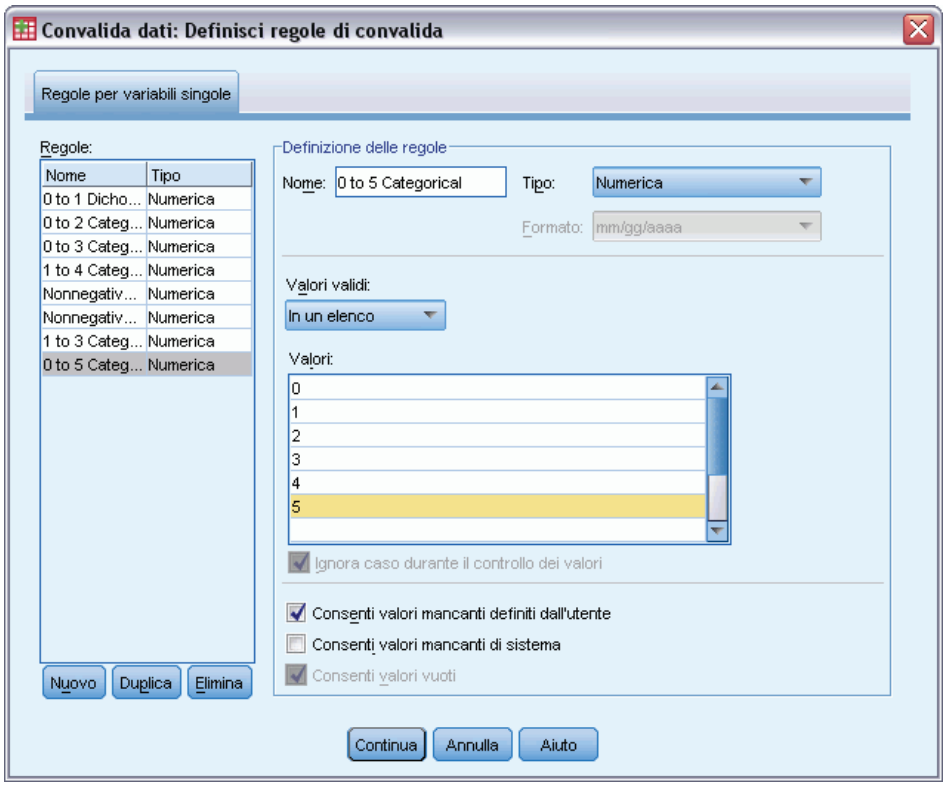
- Per definire una regola, fare clic su Nuova.

Figura 7-20
Finestra di dialogo Definisci regole di convalida, scheda Regole per variabili singole (opzione Categoriale da 1 a 3 definita)



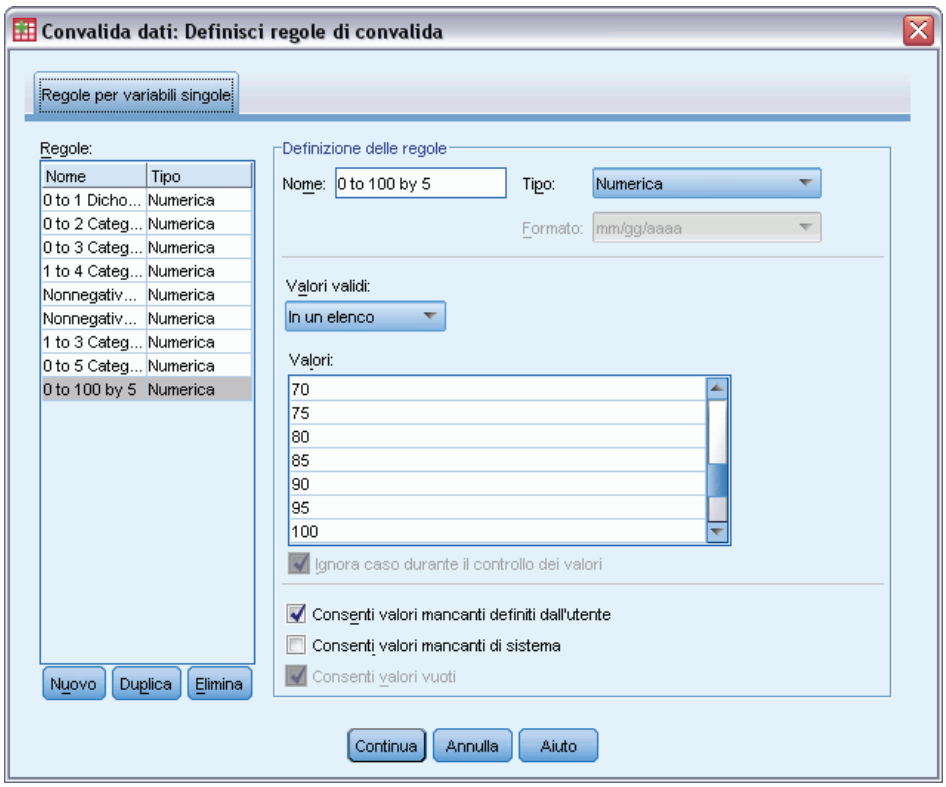
- Immettere Categoriale da 1 a 3 per il nome della regola.
- Come valori validi, selezionare In un elenco.
- Immettere 1, 2 e 3 come valori.
- Deselezionare Consenti valori mancanti di sistema.
- Per definire la regola per i punteggi Rankin, fare clic su Nuova.

Figura 7-21
Finestra di dialogo Definisci regole di convalida, scheda Regole per variabili singole (opzione Categoriale da 0 a 5 definita)



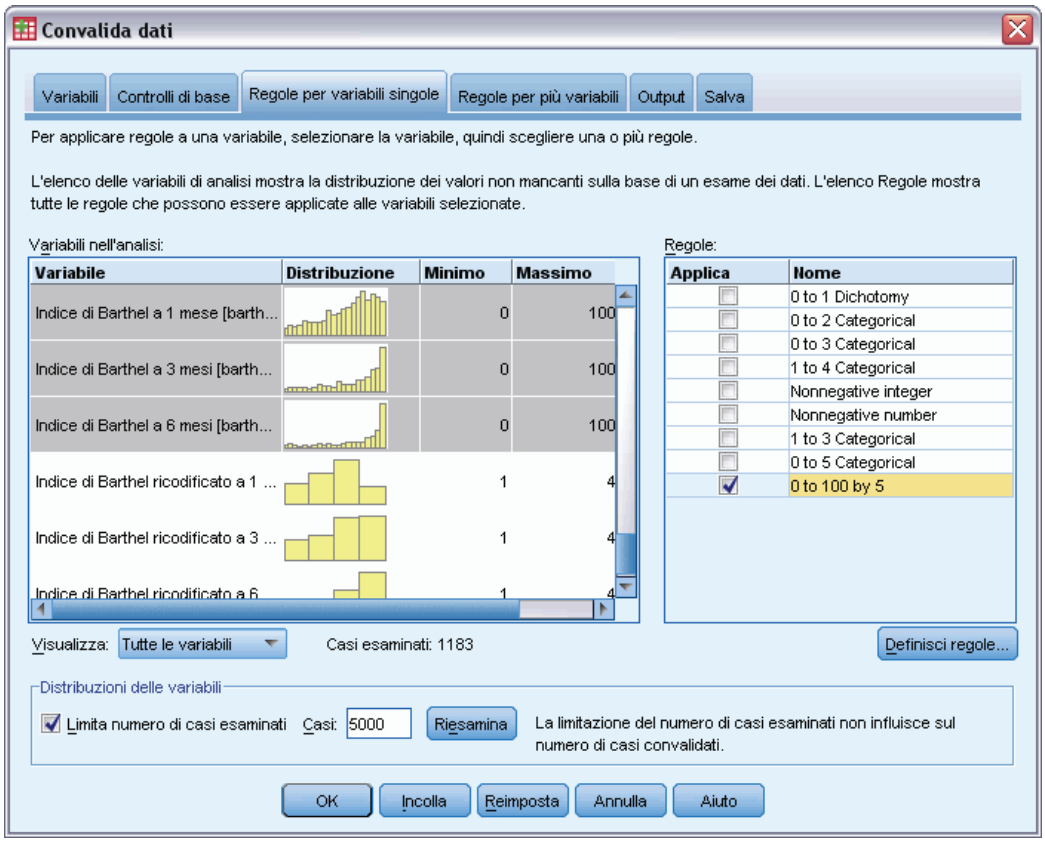
- Immettere**Categoriale da 0 a 5** per il nome della regola.
- Come valori validi, selezionare **In un elenco**.
- Immettere **0**, **1**, **2**, **3**, **4** e **5** come valori.
- Deselezionare **Consenti valori mancanti di sistema**.
- Per definire la regola per gli indici di Barthel, fare clic su **Nuova**.

Figura 7-22
Finestra di dialogo Definisci regole di convalida, scheda Regole per variabili singole (opzione da 0 a 100 per 5 definita)



- ImmettereDa 0 a 100 per 5 per il nome della regola.
- Come valori validi, selezionare In un elenco.
- Immettere 0, 5, ... e 100 come valori.
- Deselezionare Consenti valori mancanti di sistema.
- Fare clic su Continua.

Figura 7-23
Finestra di dialogo Convalida dati, scheda Regole per variabili singole (opzione da 0 a 100 per 5 definita)



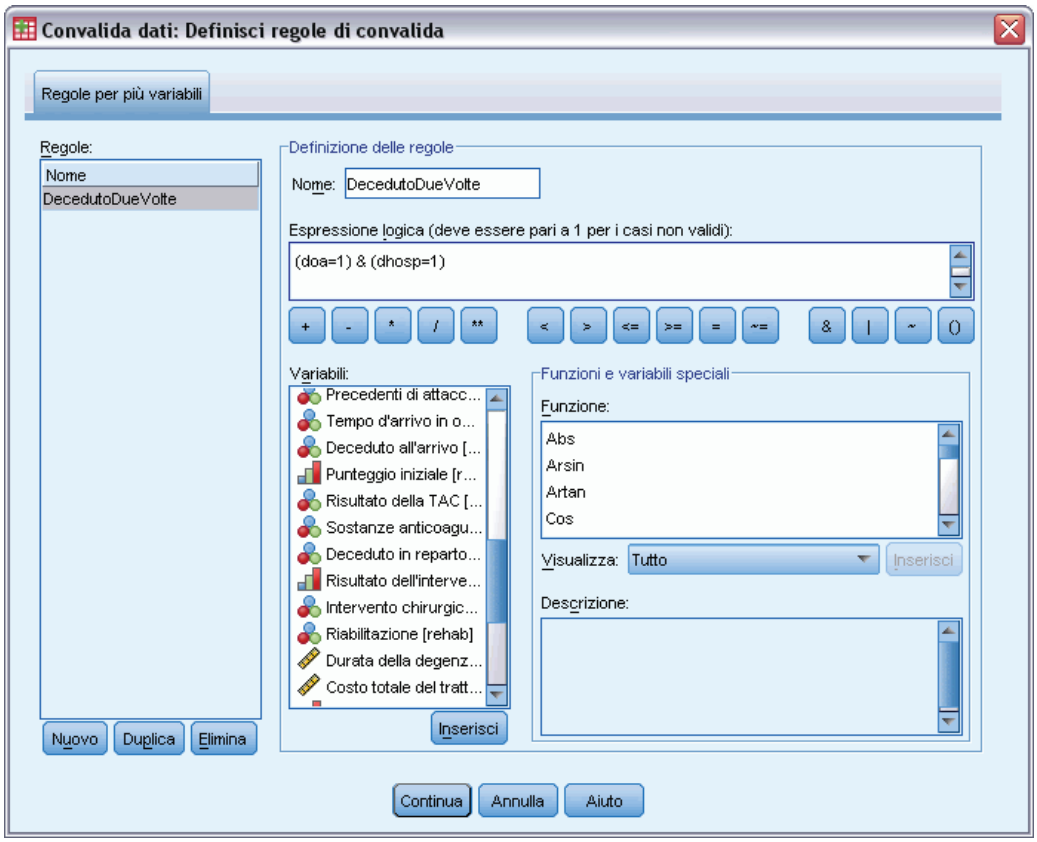
A questo punto è necessari applicare le regole definite alle variabili dell'analisi.

- Applicare Categoriale da 1 a 3 a *Dimensione ospedale*.
- Applicare Categoriale da 0 a 5 a *Punteggio Rankin iniziale* e a tutti i campi compresi tra *Punteggio Rankin dopo 1 mese* e *Punteggio Rankin dopo 6 mesi*.
- Applicare Da 0 a 100 per 5 ai campi da *Indice di Barthel dopo 1 mese* a *Indice di Barthel dopo 6 mesi*.
- Fare clic sulla scheda Regole per più variabili.

Non ci sono regole definite.

- Fare clic su Definisci regole.

Figura 7-24
Finestra di dialogo Definisci regole di convalida, scheda Regole per più variabili



Se non ci sono regole, viene automaticamente creata una nuova regola segnaposto.

- Immettere DoppioDecesso come nome della regola.
- Immettere (doa=1) & (dhosp=1) come espressione logica. In questo modo verrà restituito il valore 1 se il paziente è stato registrato sia come deceduto all'arrivo che come deceduto in ospedale.
- Fare clic su Continua.

  La nuova regola definita viene automaticamente selezionata nella schede Regola per più variabili.
- Fare clic su OK.

Regole per più variabili

Figura 7-25
Regole per più variabili

Regola	Numero di violazioni	Espressione della regola
DiedTwice	27	(doa=1) & (dhosp=1)

L'elenco riassuntivo Regole per più variabili visualizza le regole che sono state violate almeno una volta, il numero di violazioni e una descrizione di ciascuna regola violata.

Report dei casi

Figura 7-26
Report dei casi

	Violazioni delle regole di convalida		Identificatore		
Caso	Variabile singola[a]	Più variabili	hospid	patid	physid
20		Died twice	PBW	1192970826	355184
49		Died twice	NHV	8717862852	237418
129		Died twice	QWS	6901932085	215041
138		Died twice	RLD	1205005069	695521
162		Died twice	OZN	5546809538	125304
175	0 to 1 Dichotomy (1)		OZN	0333204686	883285
274	0 to 1 Dichotomy (1)		OZN	1038840465	103254
310	Nonnegative integer (1)		OZN	2090290204	883285
414		Died twice	WPA	3351107142	462020
437	0 to 1 Dichotomy (1)		WPA	2349729006	723384
447		Died twice	WPA	7163481282	519548
458		Died twice	WPA	9159094175	652070
462		Died twice	WPA	2137520354	723384
537		Died twice	SLB	5246122506	928076
544		Died twice	SLB	1605957462	506108
620		Died twice	GFG	8141858966	828754
629		Died twice	GFG	3397891610	539412
630		Died twice	GFG	3397891610	539412
639		Died twice	GFG	3962622031	327422
644		Died twice	GFG	4271782383	749432
649		Died twice	GFG	0950686750	618069
653		Died twice	GFG	0663642766	001448
722		Died twice	GFG	0418125590	877354
748		Died twice	GFG	8744721380	539412
752	Nonnegative integer (1) 0 to 1 Dichotomy (3)		GFG	4993307441	828754
868		Died twice	WWL	9714672452	237547
881		Died twice	WWL	6613279456	574275
915		Died twice	EFX	2575793702	501318
933		Died twice	IZO	2807437472	680253
1010		Died twice	BLA	5284009939	657638
1028		Died twice	BLA	8021997463	185703
1054		Died twice	ALK	0950897644	267830
1173	1 to 4 Categorical (1)		ALK	8737661990	185787

a. Il numero do variabili che ha violato la regola è mostrato dopo ciascuna regola.

Il Report dei casi visualizza ora i casi che hanno violato la regola per più variabili oltre a quelli riscontrati in precedenza e riferiti alle violazioni di regole per singole variabili. Questi casi devono essere segnalati al personale responsabile dell'immissione dati e corretti.

Riepilogo

A questo punto l'analista dispone delle informazioni necessarie per creare un report preliminare da inviare al responsabile del team di immissione dei dati.

Procedure correlate

La procedura Convalida dati è utile per controllare la qualità dei dati.

- La procedura Identifica casi anomali analizza i modelli dei dati e identifica i casi con valori significativi che differiscono dal tipo.

Capitolo
8

Preparazione automatica dati

La preparazione dei dati per l'analisi è una delle fasi più importanti in qualsiasi progetto e, in genere, una delle più lunghe. 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. L'algoritmo si può utilizzare in modo completamente **automatico**, permettendogli di scegliere e applicare le correzioni, oppure in modo **interattivo**, visualizzando in anteprima le modifiche prima che vengano apportate e accettandole o rifiutandole in funzione delle esigenze.

L'utilizzo di ADP consente di predisporre i dati per la creazione dei modelli 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.

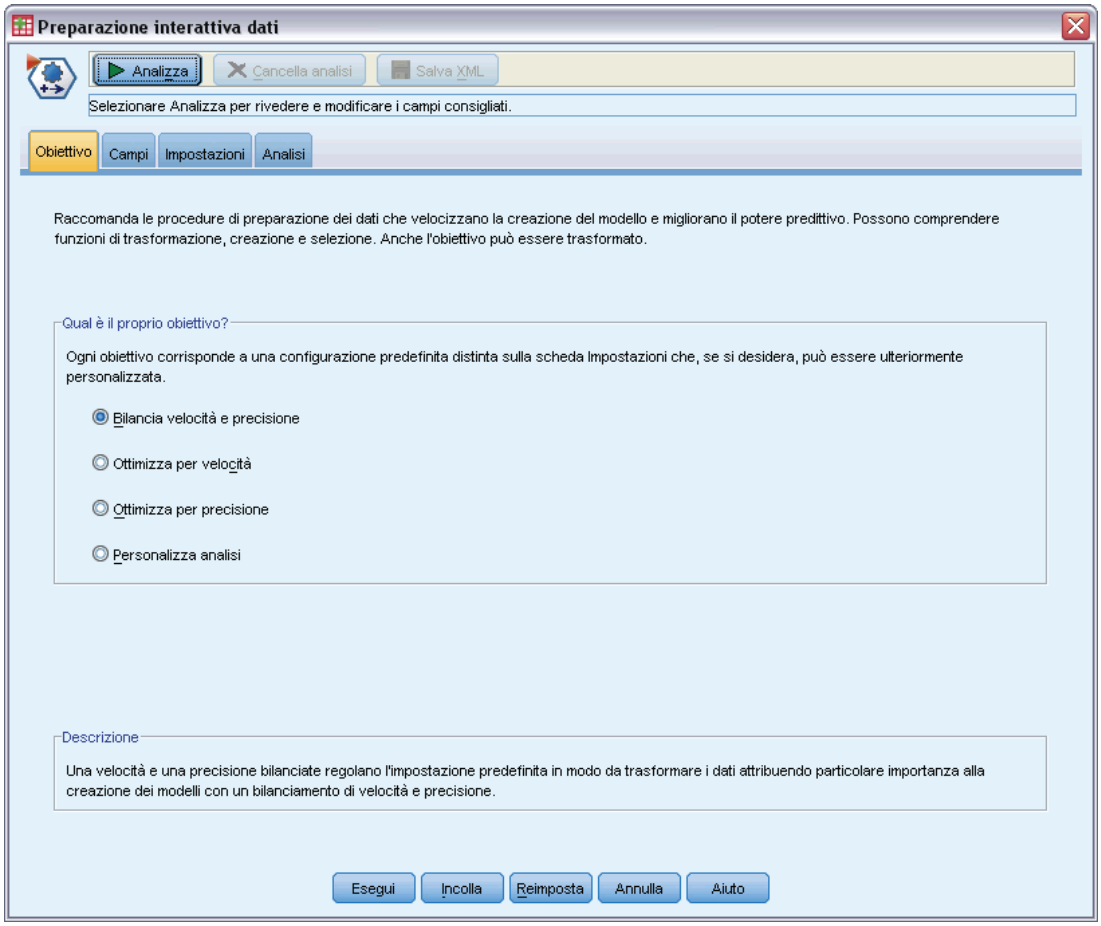
Utilizzo interattivo di Preparazione automatica dati

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. La compagnia dispone di un campione di informazioni sulle richieste di indennizzo precedenti, raccolte nel file *insurance_claims.sav*. Per ulteriori informazioni, vedere l'argomento File di esempio in l'appendice A a pag. 137. 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.

Scelta tra obiettivi

- Per eseguire Preparazione automatica dati in modo interattivo, dai menu scegliere:
  Trasforma > Prepara dati per la modellazione > Interattiva...

Figura 8-1
Scheda Obiettivo



La prima scheda chiede un obiettivo che controlli le impostazioni predefinite, ma qual è la differenza pratica tra gli obiettivi? Se si esegue la procedura utilizzando ciascuno degli obiettivi, è possibile vedere come differiscono i risultati.

- Assicurarsi che sia selezionata l'opzione Bilancia velocità e precisione, quindi fare clic su Analizza.

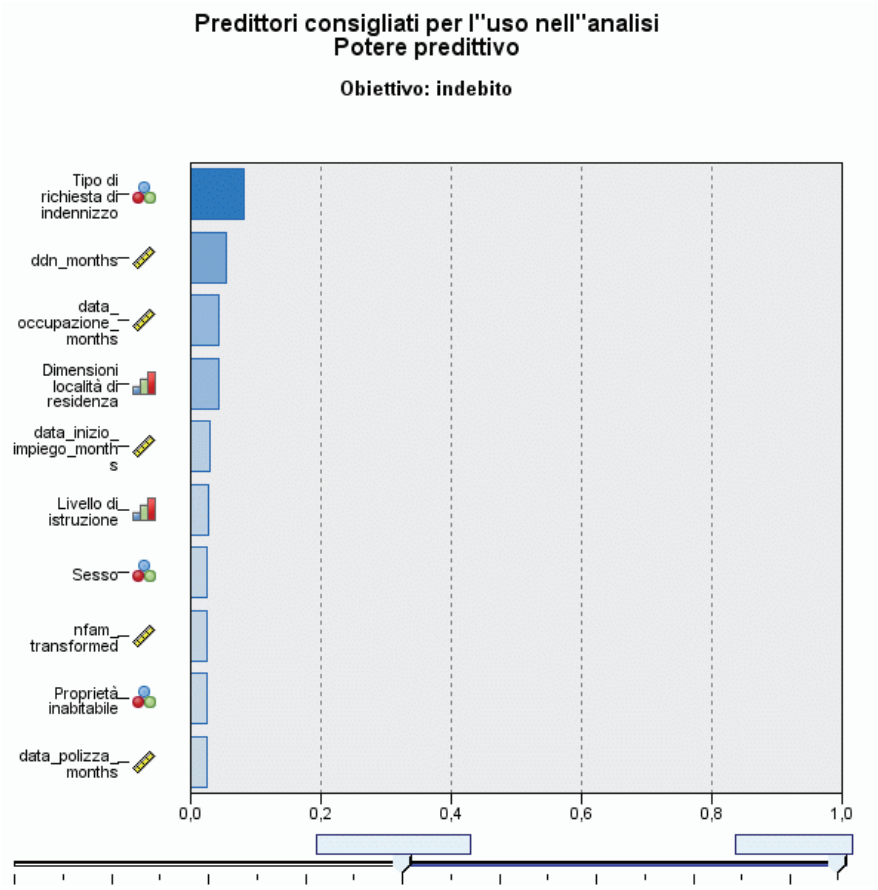
Figura 8-2
Scheda Analisi, riepilogo elaborazione campi per un obiettivo bilanciato

Riepilogo elaborazione campi

Campi		N
Obiettivo		1
Predittori		18
Predittori consigliati per l"uso nell"analisi	Totale	18
	Campi originali (non trasformati)	8
	Trasformazioni dei campi originali	5
	Derivato dalle date e ore	5
	Creato	0
Predittori non utilizzati		0

La scheda Analisi acquisisce automaticamente lo stato attivo mentre la procedura elabora i dati. La visualizzazione principale predefinita è Riepilogo elaborazione campi, che offre una panoramica di come vengono elaborati i campi dalla preparazione automatica dati. Per la creazione dei modelli sono consigliati un solo obiettivo, 18 input e 18 campi. Dei campi consigliati per la modellazione, 9 sono campi di input originali, 4 sono trasformazioni di campi di input originali e 5 sono derivati da campi di data e ora.

Figura 8-3
Scheda Analisi, potere predittivo per obiettivo bilanciato



Visualizzazione ausiliaria predefinita di Potere predittivo, che offre una rapida idea di quali, tra i campi consigliati, saranno più utili per la creazione del modello. Si noti che mentre per l'analisi sono consigliati 18 predittori, solo i primi 10 vengono mostrati nel grafico del potere predittivo per impostazione predefinita. Per mostrare un numero maggiore o minore di campi, utilizzare il comando di scorrimento sotto al grafico.

Con Bilancia velocità e precisione come obiettivo, *Tipo di richiesta di indennizzo* viene identificato come il predittore "migliore", seguito da *Numero componenti nucleo familiare* e l'età attuale del richiedente espressa in mesi (la durata calcolata dalla nascita alla data corrente).

- Fare clic su Cancella analisi, quindi sulla scheda Obiettivo.
- Selezionare Ottimizza per velocità e fare clic su Analizza.

Figura 8-4
Scheda Analisi, riepilogo elaborazione campi nel caso di ottimizzazione per velocità

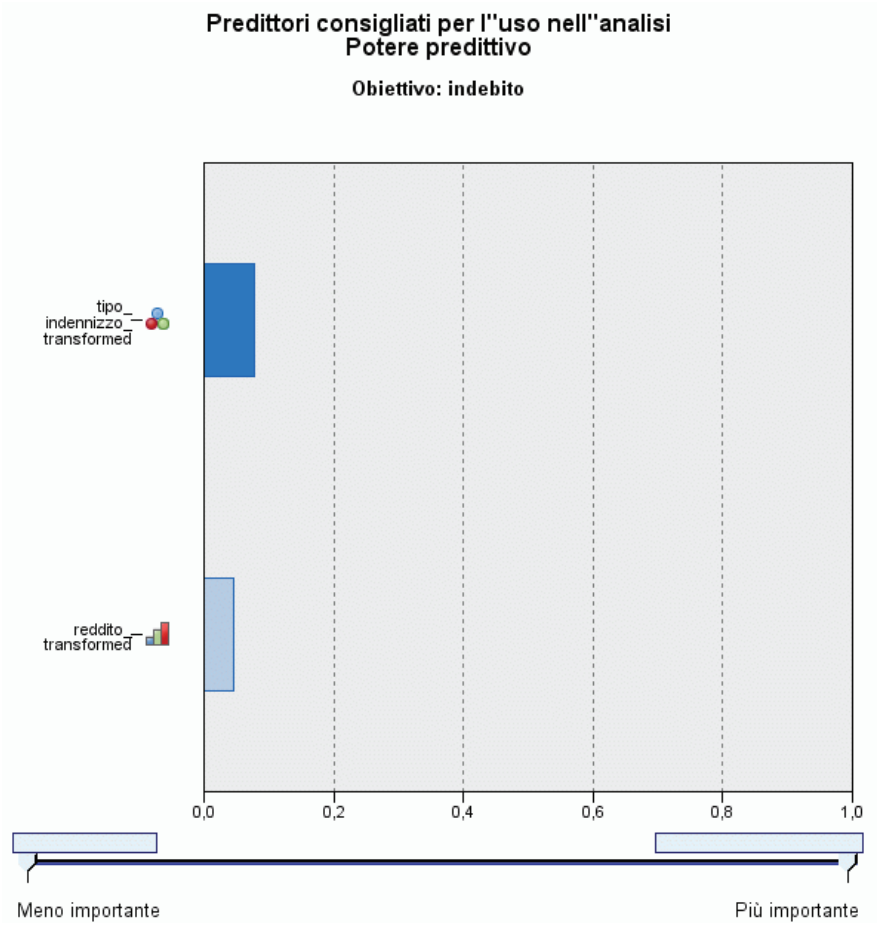
Riepilogo elaborazione campi

Campi		N
Obiettivo		1
Predittori		18
Predittori consigliati per l"uso nell"analisi	Totale	2
	Campi originali (non trasformati)	0
	Trasformazioni dei campi originali	2
	Derivato dalle date e ore	0
	Creato	0
Predittori non utilizzati		16

- Non è stato possibile creare predittori utili. I motivi più comuni sono: troppo pochi predittori continui avevano un"alta associazione all"obiettivo o tutti i predittori continui erano indipendenti.

La scheda Analisi acquisisce nuovamente e in modo automatico lo stato attivo mentre la procedura elabora i dati. In questo caso, solo 2 campi sono consigliati per la creazione del modello ed entrambi sono trasformazioni dei campi originali.

Figura 8-5
Scheda Analisi, potere predittivo in caso di ottimizzazione per velocità



Con Ottimizza per velocità come obiettivo, *tipo_indennizzo_trasformato* viene identificato come il predittore "migliore", seguito da *reddito_trasformato*.

- Fare clic su Cancella analisi, quindi sulla scheda Obiettivo.
- Selezionare Ottimizza per precisione e fare clic su Analizza.

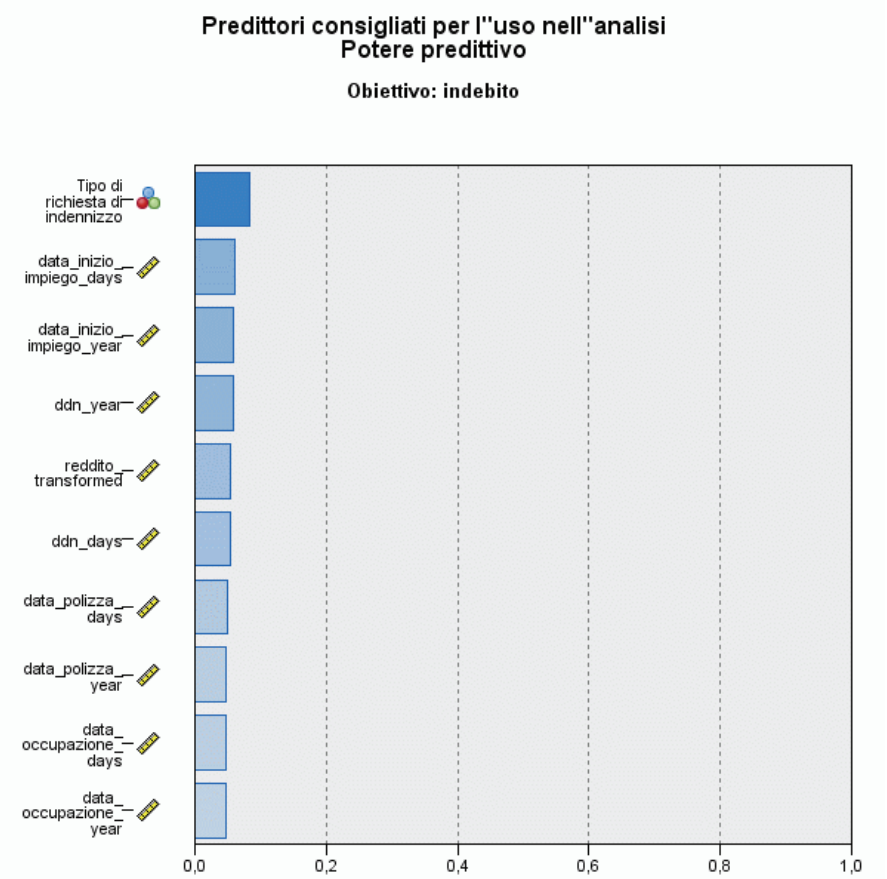
Figura 8-6
Scheda Analisi, potere predittivo in caso di ottimizzazione per precisione

Riepilogo elaborazione campi

Campi		N
Obiettivo		1
Predittori		18
Predittori consigliati per l''uso nell''analisi	Totale	32
	Campi originali (non trasformati)	8
	Trasformazioni dei campi originali	5
	Derivato dalle date e ore	19
	Creato	0
Predittori non utilizzati		0

Con Ottimizza per precisione come obiettivo, sono consigliati 32 campi per la creazione del modello, poiché viene derivato un numero maggiore di campi dalle date e dalle ore. Infatti dalle date vengono estratti i giorni, i mesi e gli anni e dalle ore, i minuti e i secondi.

Figura 8-7
Scheda Analisi, potere predittivo in caso di ottimizzazione per precisione



Tipo di richiesta di indennizzo viene identificato come il predittore "migliore", seguito dal numero di giorni da cui il richiedente occupa la posizione lavorativa più recente (la durata calcolata dalla data di inizio dell'impiego fino alla data corrente) e dall'anno in cui il richiedente ha iniziato l'impiego attuale (estratto dalla data di inizio dell'impiego).

Per riepilogare:

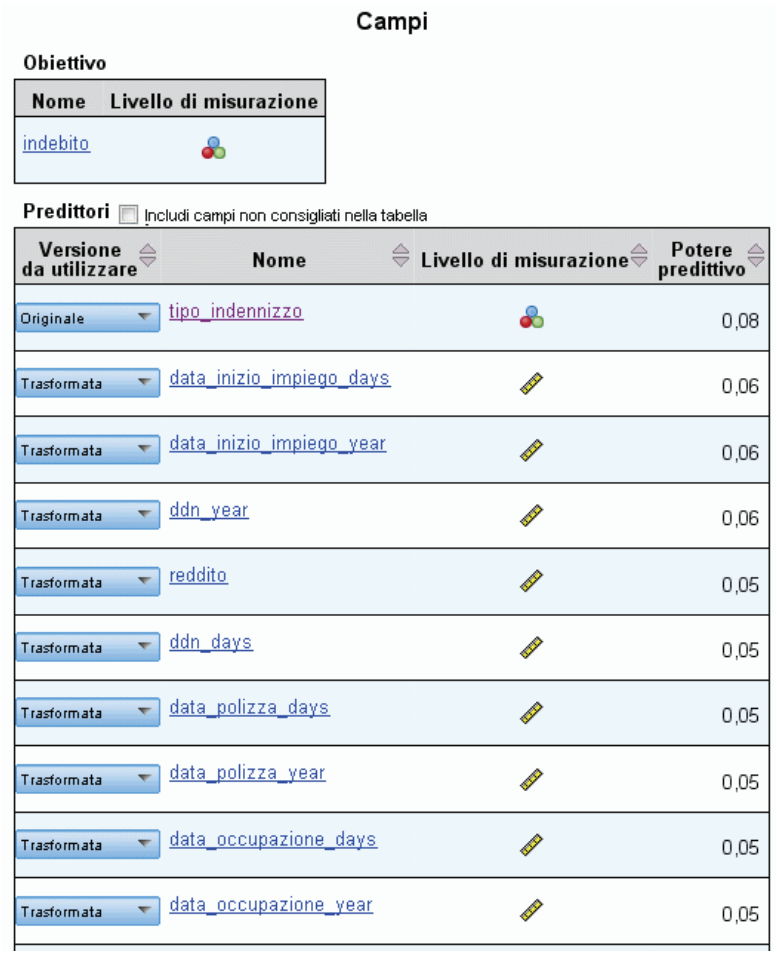
- Bilancia velocità e precisione crea campi utilizzabili nella modellazione a partire dalle date e può trasformare i campi continui come *nfam* per renderli più normalmente distribuiti.
- Ottimizza per precisione crea alcuni campi extra dalle date (oltre a verificare la presenza di valori anomali e, se l'obiettivo è continuo, può trasformarlo in modo da renderlo più normalmente distribuito).
- Ottimizza per velocità non prepara le date e non riscala i campi continui, ma unisce le categorie dei predittori categoriali e categorizza i predittori continui quando l'obiettivo è categoriale (esegue anche la selezione e creazione funzioni quando l'obiettivo è continuo).

La compagnia di assicurazioni decide di analizzare ulteriormente i risultati ottenuti da Ottimizza per precisione.

▶ Selezionare Campi dall'elenco a discesa della visualizzazione principale.

Campi e dettagli campo

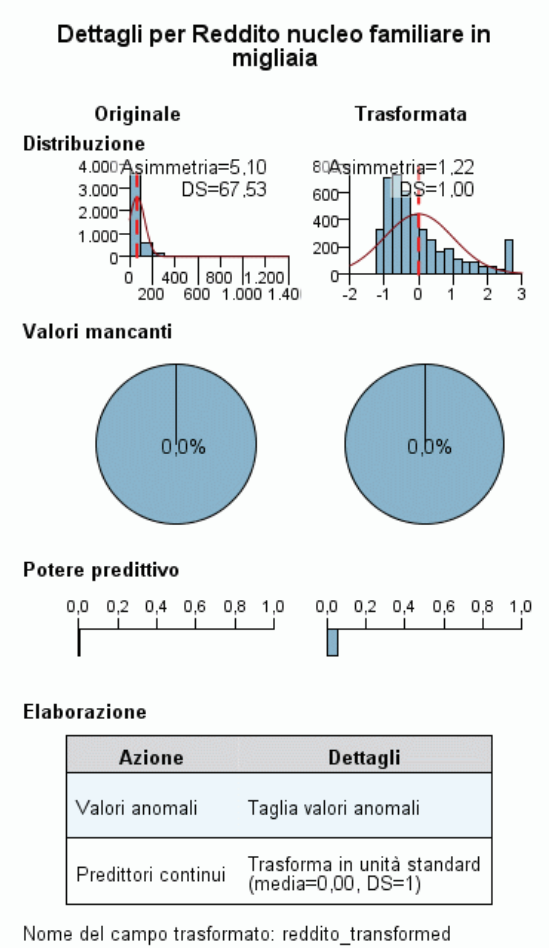
Figura 8-8
Campi



La visualizzazione Campi mostra i campi elaborati e indica se ADP ne consiglia o meno l'utilizzo nella creazione dei modelli. Facendo clic su un qualsiasi nome di campo vengono visualizzate ulteriori informazioni sul campo nella visualizzazione collegata.

▶ Fare clic su reddito.

Figura 8-9
Dettagli di campo per Reddito familiare in migliaia

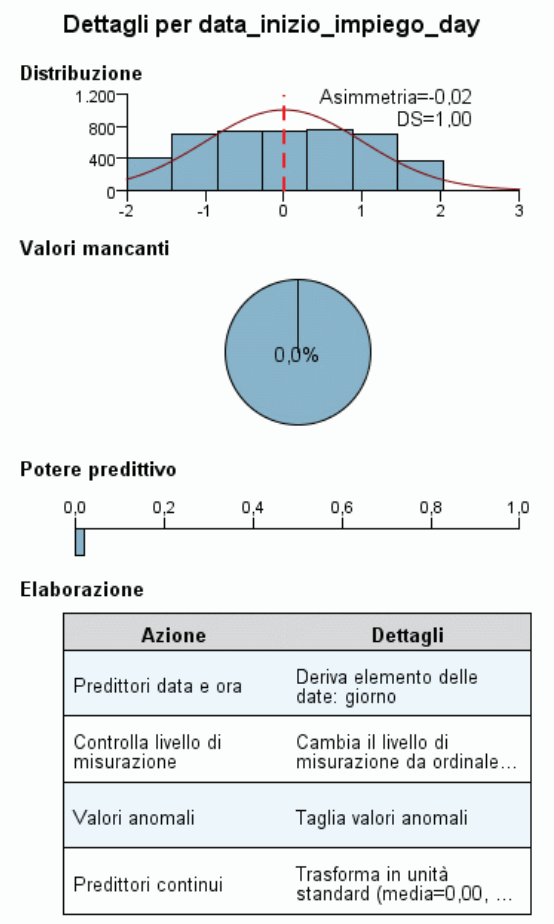


La visualizzazione Dettagli di campo mostra le distribuzioni del campo *Reddito familiare in migliaia* originale e trasformato. In base alla tabella di elaborazione, i record identificati come anomali sono stati tagliati (impostando i loro valori uguali al valore di riferimento per la determinazione dei valori anomali) e il campo è stato standardizzato in modo da avere una media pari a 0 e una deviazione standard pari a 1. La "protuberanza" all'estrema destra dell'istogramma del campo trasformato mostra che un certo numero di record, forse più di 200, è stato identificato come anomalo. Il reddito presenta una distribuzione pesantemente asimmetrica, pertanto potrebbe trattarsi di un caso in cui il valore di riferimento predefinito è troppo aggressivo nel determinare i valori anomali.

Si noti, inoltre, l'aumento di potere predittivo del campo trasformato rispetto al campo originale. Questa sembra essere una trasformazione utile.

- Nella visualizzazione Campi, fare clic su data_inizio_impiego_giorno. (Si noti che è diverso da *data_inizio_impiego_giorni*.)

Figura 8-10
Dettagli di campo per data_inizio_impiego_giorno



Il campo *data_inizio_impiego_giorno* è il giorno estratto da *Data di inizio impiego [data_inizio_impiego]*. È altamente improbabile che questo campo sia utile nell'individuazione di una richiesta di risarcimento fraudolenta, pertanto la compagnia di assicurazioni vuole rimuoverlo dalla creazione dei modelli.

Figura 8-11
Dettagli di campo per Reddito familiare in migliaia

- Nella visualizzazione Campi, selezionare Non utilizzare dall'elenco a discesa Versione da usare nella riga data_inizio_impiego_giorno. Eseguire la stessa operazione per tutti i campi con i suffissi *_giorno* e *_mese*.

- Per applicare le trasformazioni, fare clic su Esegui.

Ora l'insieme di dati è pronto per la creazione del modello, nel senso che tutti i predittori consigliati (sia nuovi sia vecchi) hanno il proprio ruolo impostato su Input, mentre il ruolo dei predittori non consigliati è impostato su Nessuno. Per creare un insieme di dati con solo i predittori consigliati, utilizzare le impostazioni Applica trasformazioni nella finestra di dialogo.

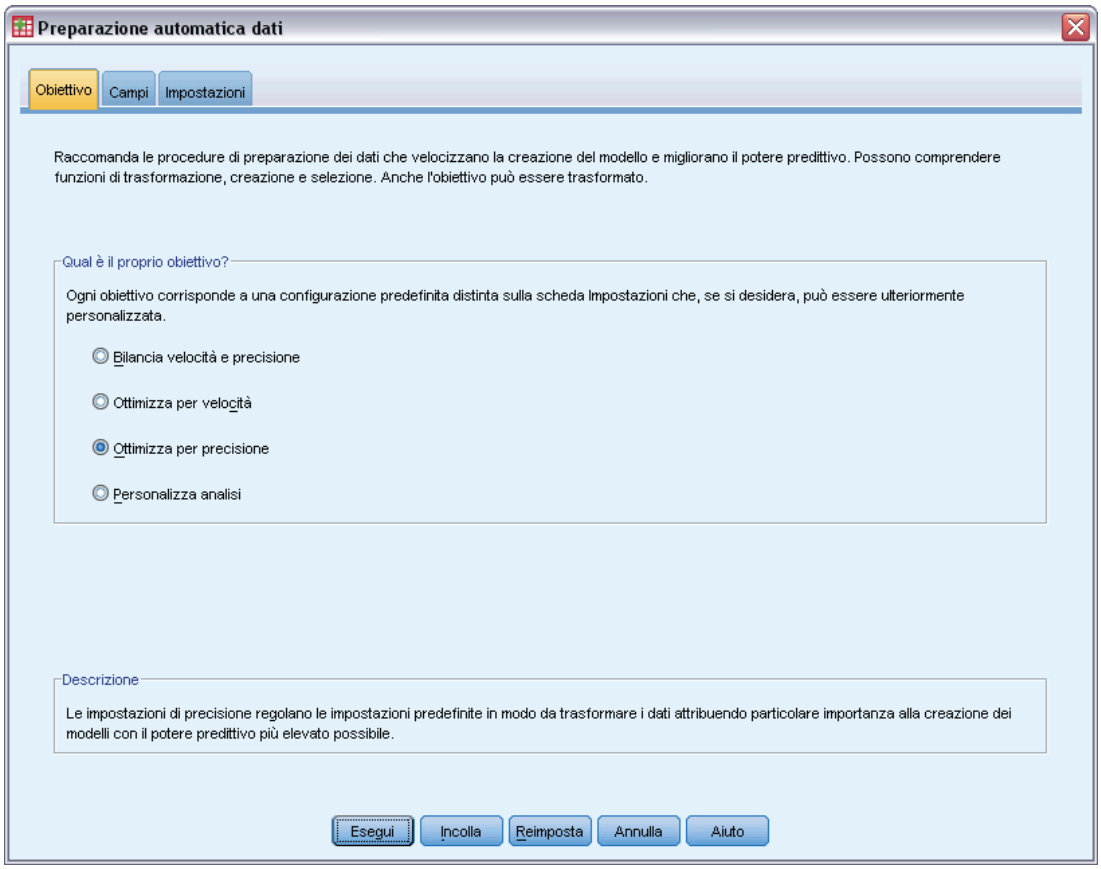
Utilizzo automatico di Preparazione automatica dati

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. Tali informazioni vengono raccolte nel file *car_sales_unprepared.sav*. Per ulteriori informazioni, vedere l'argomento File di esempio in l'appendice A a pag. 137. Utilizzare la preparazione automatica dati per preparare i dati per l'analisi. Inoltre, creare i modelli utilizzando i dati "prima" e "dopo" la preparazione in modo da poter confrontare i risultati.

Preparazione dei dati

- Per eseguire Preparazione automatica dati in modalità automatica, dai menu scegliere:
  Trasforma > Prepara dati per la modellazione > Automatica...

Figura 8-12
Scheda Obiettivo

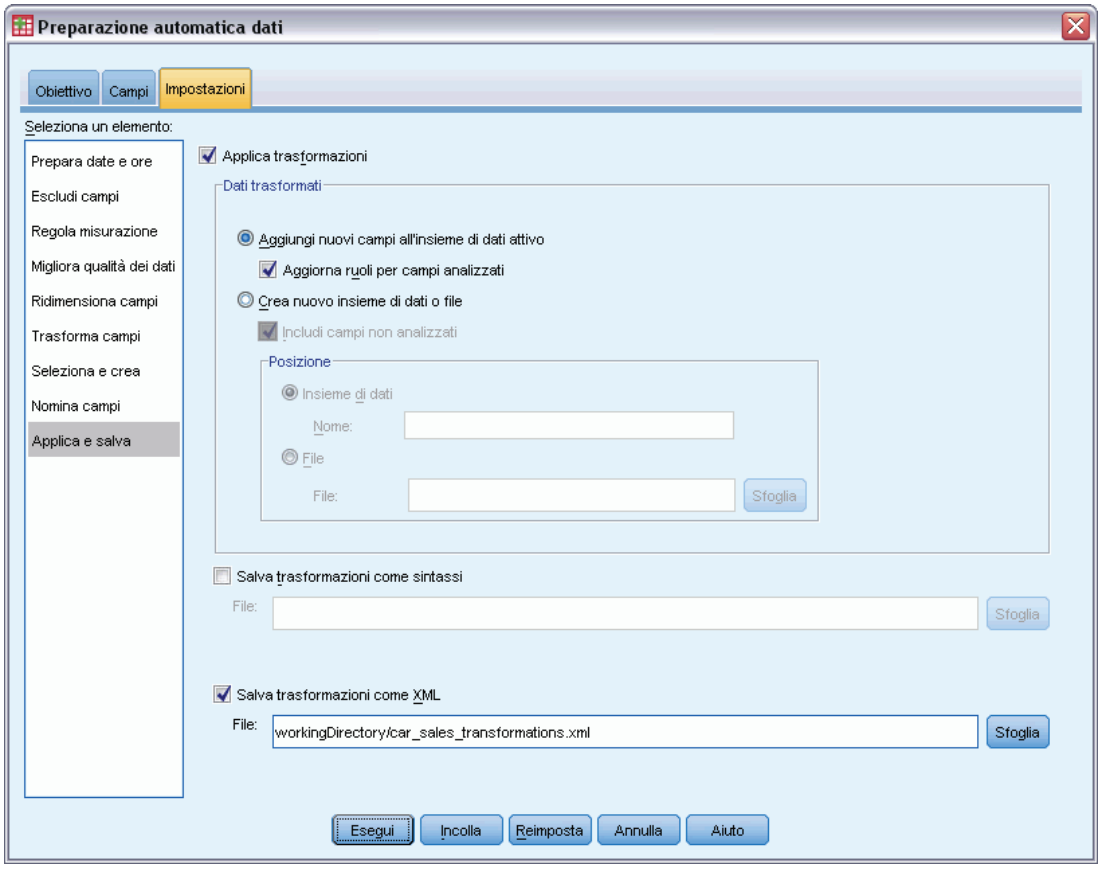


- Selezionare Ottimizza per precisione.

  Dal momento che il campo obiettivo, *Vendite in migliaia*, è continuo e può essere trasformato durante la preparazione automatica dati, è opportuno salvare le trasformazioni in un file XML per poter utilizzare la finestra di dialogo Trasforma all'indietro i punteggi e convertire i valori attesi dell'obiettivo trasformato alla scala originale.

- Fare clic sulla scheda Impostazioni, quindi sulle impostazioni Applica e salva.

Figura 8-13
Impostazioni Applica e salva



- Selezionare Salva trasformazioni come XML e fare clic su Sfoglia per navigare al file Directorydilavoro/car_sales_transformations.xml, sostituendo Directorydilavoro con il percorso nel quale si desidera salvare il file.
- Fare clic su Esegui.

Tali selezioni generano la sintassi seguente:

```
*Automatic Data Preparation.
ADP
  /FIELDS TARGET=sales INPUT=resale type price engine_s horsepow wheelbas width length
    curb_wgt fuel_cap mpg
  /PREPDATETIME DATEDURATION=YES(REFERENCE=YMD('2009-06-04') UNIT=AUTO)
    TIMEDURATION=YES(REFERENCE=HMS('08:43:35') UNIT=AUTO) EXTRACTYEAR=YES(SUFFIX='_year')
    EXTRACTMONTH=YES(SUFFIX='_month') EXTRACTDAY=YES(SUFFIX='_day')
    EXTRACTHOUR=YES(SUFFIX='_hour') EXTRACTMINUTE=YES(SUFFIX='_minute')
    EXTRACTSECOND=YES(SUFFIX='_second')
   /SCREENING PCTMISSING=YES(MAXPCT=50) UNIQUECAT=YES(MAXCAT=100) SINGLECAT=NO
  /ADJUSTLEVEL INPUT=YES TARGET=YES MAXVALORDINAL=10 MINVALCONTINUOUS=5
  /OUTLIERHANDLING INPUT=YES TARGET=NO CUTOFF=SD(3) REPLACEWITH=CUTOFFVALUE
  /REPLACEMISSING INPUT=YES TARGET=NO
  /REORDERNOMINAL INPUT=YES TARGET=NO
  /RESCALE INPUT=ZSCORE(MEAN=0 SD=1) TARGET=BOXCOX(MEAN=0 SD=1)
  /TRANSFORM MERGESUPERVISED=NO MERGEUNSUPERVISED=NO BINNING=NONE SELECTION=NO
    CONSTRUCTION=NO
```

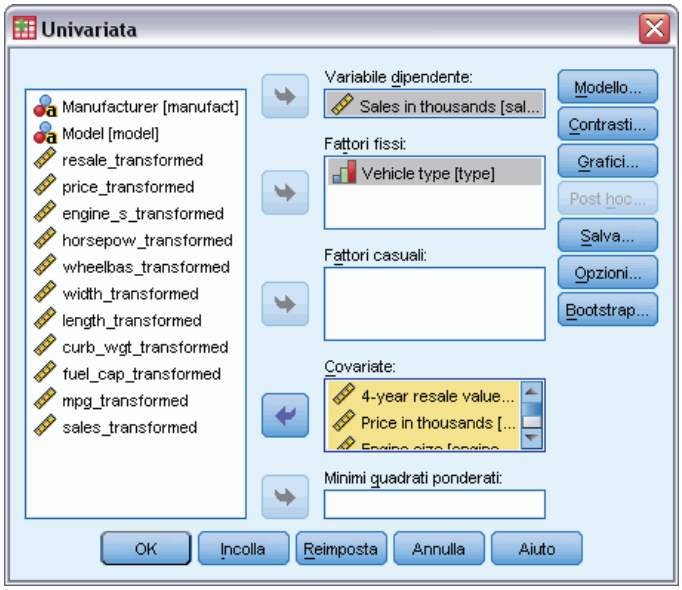
```
  /CRITERIA SUFFIX(TARGET='_transformed' INPUT='_transformed')
  /OUTFILE PREPXML='/workingDirectory/car_sales_transformations.xml'.
TMS IMPORT
  /INFILE TRANSFORMATIONS='/workingDirectory/car_sales_transformations.xml'
    MODE=FORWARD (ROLES=UPDATE)
  /SAVE TRANSFORMED=YES.
EXECUTE.
```

- Il comando `ADP` prepara il campo obiettivo *vendite* e i campi di input *rivendita* mediante *mpg*.
- Il sottocomando `PREPDATETIME` viene specificato ma non utilizzato perché nessuno dei campi è di tipo data o ora.
- Il sottocomando `ADJUSTLEVEL` riformula i campi ordinali con più di 10 valori come continui e i campi continui con meno di 5 valori come ordinali.
- Il sottocomando `OUTLIERHANDLING` sostituisce i valori degli input continui (non dell'obiettivo) con più di 3 deviazioni standard rispetto alla media con il valore corrispondente a 3 deviazioni standard rispetto alla media.
- Il sottocomando `REPLACEMISSING` sostituisce i valori degli input (non dell'obiettivo) mancanti.
- Il sottocomando `REORDERNOMINAL` ricodifica i valori degli input nominali da valore meno ricorrente a valore più ricorrente.
- Il sottocomando `RESCALE` standardizza gli input continui assegnando loro media 0 e deviazione standard 1 mediante una trasformazione punteggio Z e standardizza l'obiettivo continuo assegnandogli media 0 e deviazione standard 1 mediante una trasformazione di Box-Cox.
- Il sottocomando `TRANSFORM` disattiva tutte le operazioni predefinite specificate da questo sottocomando.
- Il sottocomando `CRITERIA` specifica i suffissi predefiniti per le trasformazioni dell'obiettivo e degli input.
- Il sottocomando `OUTFILE` indica che le trasformazioni devono essere salvate in */Directorydilavoro/car_sales_transformations.xml*, dove */Directorydilavoro* è il percorso n cui si desidera salvare il file *car_sales_transformations.xml*.
- Il comando `TMS IMPORT` legge le trasformazioni in *car_sales_transformations.xml* e le applica all'insieme di dati attivo, aggiornando i ruoli dei campi esistenti che vengono trasformati.
- Il comando `EXECUTE` avvia l'elaborazione delle trasformazioni. Quando lo si utilizza nell'ambito di un flusso di sintassi più lungo, può essere possibile eliminare il comando `EXECUTE` per abbreviare il tempo di elaborazione.

Creazione di un modello su dati non preparati

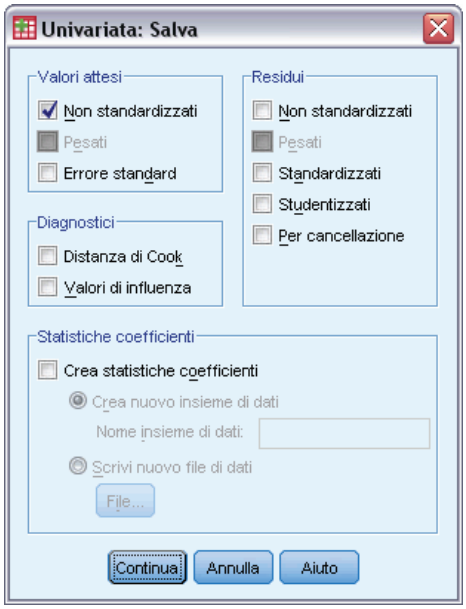
▶ Per creare un modello su dati non preparati, dai menu scegliere:
Analizza > Modello lineare generalizzato > Univariata...

Figura 8-14
Finestra di dialogo GLM univariato



- Selezionare *Vendite in migliaia [vendite]* come variabile dipendente.
- Selezionare *Tipo veicolo [tipo]* come un fattore fisso.
- Selezionare da *Valore di rivendita a 4 anni [rivendita]* a *Rendimento carburante [mpg]* come covariate.
- Scegliere Salva.

Figura 8-15
Finestra di dialogo Salva



- Selezionare Non standardizzati nel gruppo Valori attesi.
- Fare clic su Continua.
- Nella finestra di dialogo GLM univariato fare clic su OK.

Tali selezioni generano la sintassi seguente:

```
UNIANOVA sales BY type WITH resale price engine_s horsepow wheelbas width length
    curb_wgt fuel_cap mpg
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /SAVE=PRED
  /CRITERIA=ALPHA(0.05)
  /DESIGN=resale price engine_s horsepow wheelbas width length curb_wgt fuel_cap
    mpg type.
```

Figura 8-16
Effetti tra soggetti per un modello basato su dati non preparati

Variabile dipendente:Sales in thousands

Sorgente	Somma dei quadrati Tipo III	df	Media dei quadrati	F	Sig.
Modello corretto	226123.658[a]	11	20556.696	5.050	.000
Intercetta	12227.688	1	12227.688	3.004	.086
resale	50.702	1	50.702	.012	.911
price	471.630	1	471.630	.116	.734
engine_s	19872.712	1	19872.712	4.882	.029
horsepow	9644.486	1	9644.486	2.369	.127
wheelbas	29824.272	1	29824.272	7.327	.008
width	263.465	1	263.465	.065	.800
length	1374.525	1	1374.525	.338	.562
curb_wgt	32762.692	1	32762.692	8.049	.005
fuel_cap	1124.237	1	1124.237	.276	.600
mpg	337.585	1	337.585	.083	.774
type	17668.779	1	17668.779	4.341	.040
Errore	427402.183	105	4070.497		
Totale	1062354.955	117			
Totale corretto	653525.841	116			

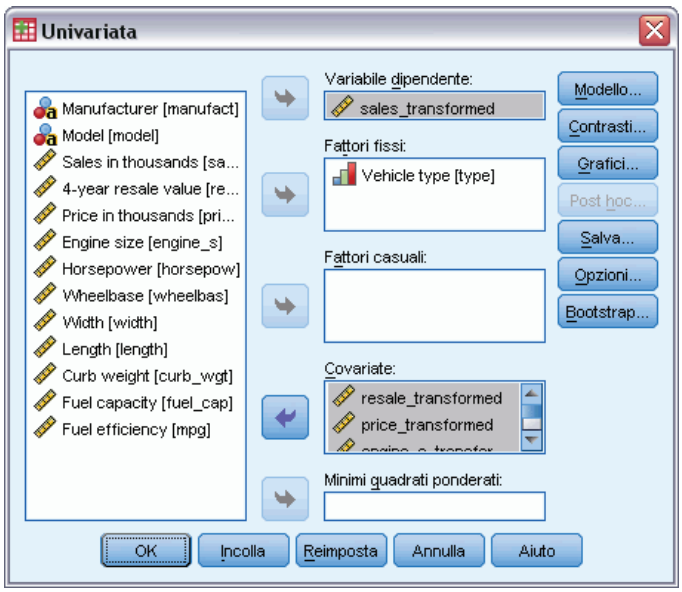
a. R quadrato = .346 (R quadrato corretto = .277)

L'output GLM univariato predefinito include gli effetti tra soggetti, che è un'analisi della tabella della varianza. Ciascun termine nel modello, più il modello nella sua interezza, viene testato per la capacità di rappresentare la variazione nella variabile dipendente. Si noti che le etichette variabili non vengono visualizzate in questa tabella.

I predittori mostrano livelli diversi di significatività; i predittori con valori di significatività inferiori a 0,05 sono solitamente considerati utili per il modello.

Creazione di un modello su dati preparati

Figura 8-17
Finestra di dialogo GLM univariato



- Per creare il modello sui dati preparati, richiamare la finestra di dialogo GLM univariato.
- Deselezionare *Vendite in migliaia [vendite]* e selezionare *vendite_trasformate* come la variabile dipendente.
- Deselezionare da *Valore di rivendita a 4 anni [rivendita]* a *Rendimento carburante [mpg]* e selezionare da *rivendita_trasformata* a *mpg_trasformate* come covariate.
- Fare clic su OK.

Tali selezioni generano la sintassi seguente:

```
UNIANOVA sales_transformed BY type WITH resale_transformed price_transformed
    engine_s_transformed horsepow_transformed wheelbas_transformed width_transformed
    length_transformed curb_wgt_transformed fuel_cap_transformed mpg_transformed
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /SAVE=PRED
  /CRITERIA=ALPHA(0.05)
  /DESIGN=resale_transformed price_transformed engine_s_transformed horsepow_transformed
    wheelbas_transformed width_transformed length_transformed curb_wgt_transformed
    fuel_cap_transformed mpg_transformed type.
```

Figura 8-18
Effetti tra soggetti per un modello basato su dati preparati

Variabile dipendente:sales_transformed

Sorgente	Somma dei quadrati Tipo III	df	Media dei quadrati	F	Sig.
Modello corretto	79.327[a]	11	7.212	13.638	.000
Intercetta	2.436	1	2.436	4.606	.034
resale_transformed	.954	1	.954	1.804	.181
price_transformed	9.271	1	9.271	17.533	.000
engine_s_transformed	2.885	1	2.885	5.456	.021
horsepow_transformed	.034	1	.034	.064	.801
wheelbas_transformed	1.213	1	1.213	2.293	.132
width_transformed	.037	1	.037	.071	.791
length_transformed	.265	1	.265	.501	.480
curb_wgt_transformed	.103	1	.103	.194	.660
fuel_cap_transformed	.132	1	.132	.249	.618
mpg_transformed	3.390	1	3.390	6.411	.012
type	4.007	1	4.007	7.579	.007
Errore	76.673	145	.529		
Totale	156.000	157			
Totale corretto	156.000	156			

a. R quadrato = .509 (R quadrato corretto = .471)

Si possono notare alcune differenze interessanti tra gli effetti tra soggetti per il modello creato sui dati non preparati e il modello creato sui dati preparati. Innanzitutto, si noti che i gradi di libertà totali sono aumentati. Ciò è dovuto al fatto che i valori mancanti sono stati sostituiti dai valori assegnati durante la preparazione automatica dati, pertanto i record rimossi in modo listwise dal primo modello sono disponibili per il secondo. Più interessante ancora è, forse, il fatto che la significatività di certi predittori è cambiata. Mentre entrambi i modelli concordano sul fatto che le dimensioni del motore [*dim_motore*] e il tipo di veicolo [*tipo*] sono utili per il modello, l'interasse [*interasse*] e il peso in ordine di marcia [*peso_ordinemarc*] non sono più significativi, e il prezzo del veicolo [*prezzo_trasformato*] e il rendimento carburante [*mpg_trasformate*] sono ora significativi.

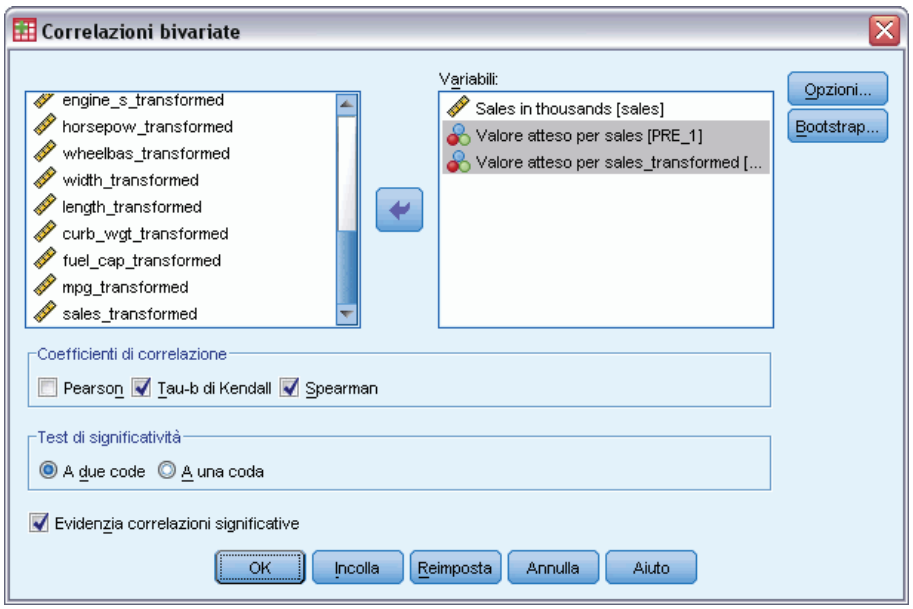
Qual è il motivo di questo cambiamento? Le vendite hanno una distribuzione asimmetrica, quindi potrebbe essere che l'interasse e il peso in ordine di marcia avessero alcuni record influenti che hanno perso la loro importanza una volta che le vendite sono state trasformate. Un'altra possibilità è che i casi extra, disponibili in seguito alla sostituzione dei valori mancanti, abbiano cambiato la significatività statistica di queste variabili. In ogni caso, si rende necessaria un'indagine più approfondita che non tratteremo in questa sede.

Si noti che R quadrato è più alto nel modello creato sui dati preparati, ma, poiché le vendite sono state trasformate, potrebbe non essere il parametro migliore per confrontare la prestazione di ciascun modello. È possibile, invece, calcolare le correlazioni non parametriche tra i valori osservati e i due insiemi di valori attesi.

Confronto tra valori attesi

- Per ottenere le correlazioni dei valori attesi dai due modelli, dai menu scegliere:
  Analizza > Correlazione > Bivariata...

Figura 8-19
Finestra di dialogo Correlazioni bivariate



- Selezionare *Vendite in migliaia [vendite]*, *Valore atteso per vendite [ATT_1]* e *Valori attesi per vendite_trasformate [ATT_2]* come variabili di analisi.
- Deselezionare Pearson e selezionare Tau-b di Kendall e Spearman nel gruppo Coefficienti di correlazione.

  Si noti che *Valori attesi per vendite_trasformate [ATT_2]* può essere utilizzato per calcolare le correlazioni non parametriche senza doverle trasformare all'indietro alla scala originale, dal momento che la trasformazione all'indietro non cambia l'ordine di classificazione dei valori attesi.
- Fare clic su OK.

  Tali selezioni generano la sintassi seguente:

```
NONPAR CORR
  /VARIABLES=sales PRE_1 PRE_2
  /PRINT=BOTH TWOTAIL NOSIG
  /MISSING=PAIRWISE.
```

Figura 8-20
Correlazioni non parametriche

			Sales in thousands	Valore atteso per sales	Valore atteso per sales_ transformed
Tau_b di Kendall	Sales in thousands	Coefficiente di correlazione	1.000	.376**	.484**
		Sig. (2-code)	.	.000	.000
		N	157	117	157
	Valore atteso per sales	Coefficiente di correlazione	.376**	1.000	.655**
		Sig. (2-code)	.000	.	.000
		N	117	117	117
	Valore atteso per sales_transformed	Coefficiente di correlazione	.484**	.655**	1.000
		Sig. (2-code)	.000	.000	.
		N	157	117	157
Rho di Spearman	Sales in thousands	Coefficiente di correlazione	1.000	.530**	.666**
		Sig. (2-code)	.	.000	.000
		N	157	117	157
	Valore atteso per sales	Coefficiente di correlazione	.530**	1.000	.831**
		Sig. (2-code)	.000	.	.000
		N	117	117	117
	Valore atteso per sales_transformed	Coefficiente di correlazione	.666**	.831**	1.000
		Sig. (2-code)	.000	.000	.
		N	157	117	157

**. La correlazione è significativa al livello 0,01 (2-code).

La prima colonna mostra che i valori attesi per il modello creato utilizzando i dati preparati sono correlati in modo più forte con i valori osservati sia dal parametro Tau-b di Kendall sia dal parametro rho di Spearman. Ciò suggerisce che l'esecuzione della preparazione automatica dati ha migliorato il modello.

Trasformazione all'indietro dei valori attesi

- I dati preparati includono una trasformazione delle vendite, pertanto i valori attesi da questo modello non sono direttamente utilizzabili come punteggi. Per trasformare i valori attesi nella scala originale, dai menu scegliere:
  Trasforma > Prepara dati per la modellazione > Trasforma all'indietro i punteggi...

Figura 8-21
Finestra di dialogo Trasforma all'indietro i punteggi

- Selezionare *Valore atteso per vendite_trasformate [ATT_2]* come campo da trasformare all'indietro.
- Digitare _trasformato_all_indietro come suffisso per il nuovo campo.
- Digitare Directorydilavoro\car_sales_transformations.xml, sostituendo Directorydilavoro, come posizione del file XML che contiene le trasformazioni, con il percorso al file.
- Fare clic su OK.

Tali selezioni generano la sintassi seguente:

```
TMS IMPORT
  /INFILE TRANSFORMATIONS='workingDirectory/car_sales_transformations.xml'
    MODE=BACK (PREDICTED=PRE_2 SUFFIX='_backtransformed').
EXECUTE.
```

- Il comando `TMS IMPORT` legge le trasformazioni in *car_sales_transformations.xml* e applica la trasformazione all'indietro a *PRE_2*.
- Il nuovo campo contenente i valori trasformati all'indietro viene denominato *PRE_2_backtransformed*.
- Il comando `EXECUTE` avvia l'elaborazione delle trasformazioni. Quando lo si utilizza nell'ambito di un flusso di sintassi più lungo, può essere possibile eliminare il comando `EXECUTE` per abbreviare il tempo di elaborazione.

Riepilogo

L'utilizzo della preparazione automatica dati consente di ottenere rapidamente le trasformazioni dei dati che possono migliorare il modello. Se l'obiettivo è trasformato, è possibile salvare le trasformazioni in un file XML e utilizzare la finestra di dialogo Trasforma all'indietro i punteggi per convertire i valori attesi per l'obiettivo trasformato alla scala originale.

Capitolo
9

Identifica casi anomali

La procedura Rilevamento anomalie ricerca i casi insoliti in base agli scostamenti dalle norme dei gruppi cluster corrispondenti. La procedura permette di rilevare rapidamente i casi insoliti per il controllo dei dati nella fase di analisi esplorativa dei dati, prima di effettuare l'analisi inferenziale dei dati. Questo algoritmo è progettato per il rilevamento di casi anomali generici. In altre parole la definizione di casi anomali non si riferisce a un'applicazione specifica, come il rilevamento degli schemi di pagamento anomali nell'industria sanitaria o il rilevamento di casi di riciclaggio nell'industria finanziaria in cui un'anomalia può essere definita in modo specifico.

Algoritmo Identifica casi anomali

Questo algoritmo è diviso in tre fasi:

Creazione di modelli. La procedura crea un modello di cluster che mostra i raggruppamenti naturali (o cluster) del file dati che non sarebbero altrimenti visibili. I cluster sono basati su un insieme di variabili di input. Il modello di cluster risultante e i dati statistici necessari per calcolare le norme del gruppo di cluster vengono salvati per riferimento.

Assegnazione di punteggi. Il modello viene applicato a ciascun caso e usato per identificare il gruppo di cluster corrispondente. Vengono creati anche alcuni indici per ciascun caso per consentire la misurazione delle differenze del caso rispetto al gruppo di cluster. Tutti i casi vengono ordinati in base ai valori degli indici di anomalia. La sezione superiore dell'elenco dei casi viene identificata come l'insieme delle anomalie.

Creazione di motivi. Le variabili di ciascun caso anomalo vengono ordinate in base agli indici di deviazione delle variabili corrispondenti. Le variabili superiori, i relativi valori e i corrispondenti valori di norma vengono presentati come motivi che spiegano perché il caso è stato identificato come anomalo.

Identificazione di casi anomali in un database medico

Ad un analista viene chiesto di preparare dei modelli predittivi per valutare i risultati dei trattamenti ai pazienti vittime di infarti cardiaci. L'analista è preoccupato della qualità dei dati perché questi modelli sono sensibili alle osservazioni anomale. Alcune di queste osservazioni

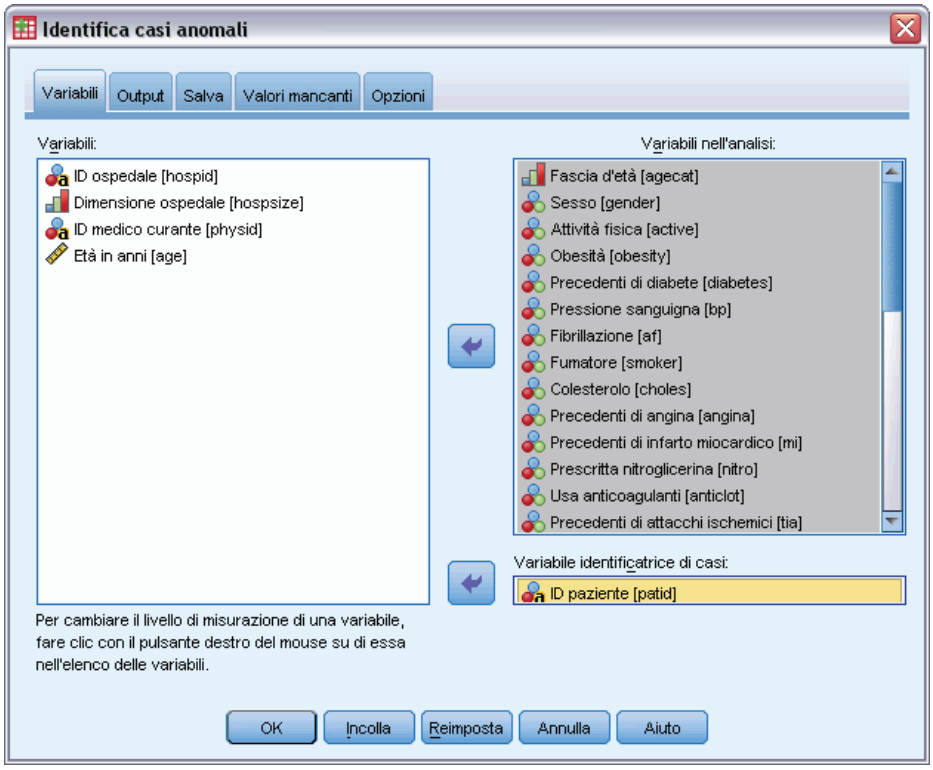
anomale rappresentano di fatto casi univoci e non sono indicati per la previsione, mentre altre sono dovute ad errori di inserimento dati, ovvero a casi in cui i valori sono tecnicamente "corretti" che non vengono quindi rilevati dalle procedure di convalida dati.

Tali informazioni vengono raccolte nel file *stroke_valid.sav*. Per ulteriori informazioni, vedere l'argomento File di esempio in l'appendice A a pag. 137. Utilizzare la procedura Identifica casi anomali per pulire il file di dati. La sintassi utilizzabile per riprodurre queste analisi è contenuta in *detectanomaly_stroke.sps*.

Esecuzione dell'analisi

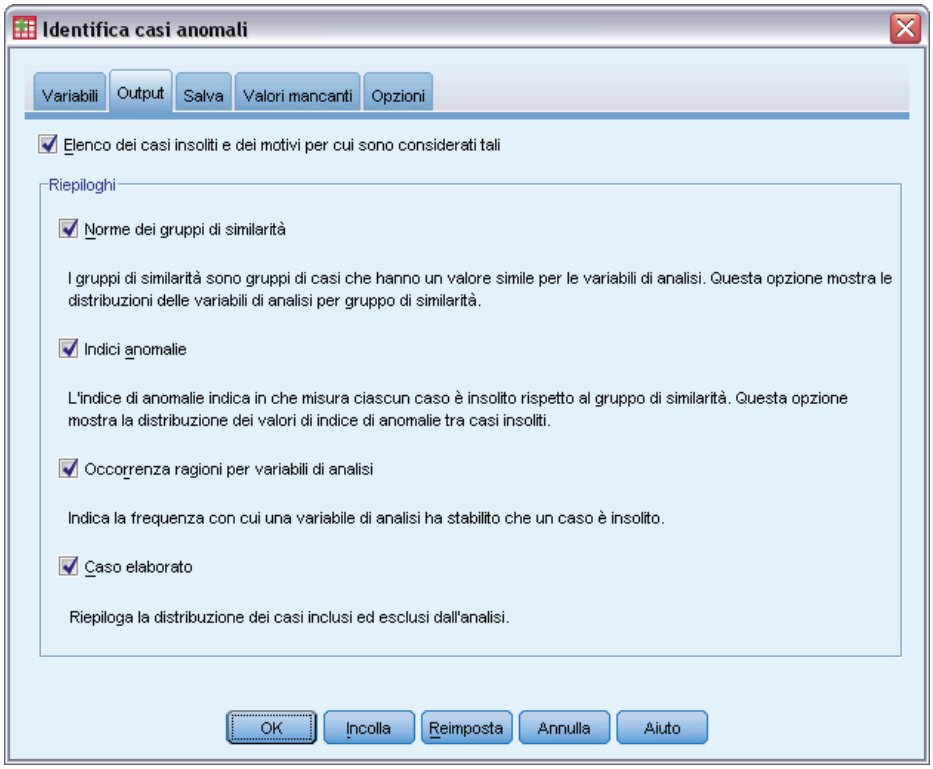
- Per identificare i casi anomali, selezionare dai menu:
  Dati > Identifica casi anomali...

Figura 9-1
Scheda Variabili della finestra di dialogo Identifica casi anomali



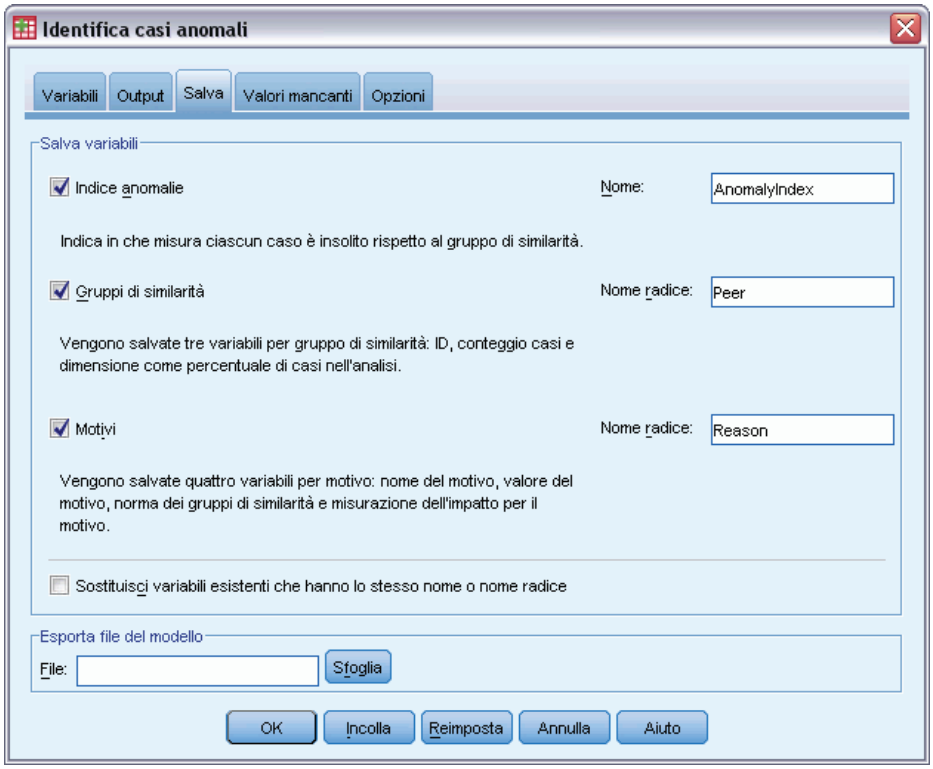
- Selezionare da *Categoria età* a *Ictus tra 3 e 6 mesi* come variabili dell'analisi.
- Selezionare *ID paziente* come variabile di identificazione dei casi.
- Fare clic sulla scheda Output.

Figura 9-2
Scheda Output della finestra di dialogo Identifica casi anomali



- Selezionare Norme dei gruppi equivalenti, Indici di anomalia, Occorrenza motivi per variabile di analisi e Casi elaborati.
- Fare clic sulla scheda Salva.

Figura 9-3
Scheda Salva della finestra di dialogo Identifica casi anomali

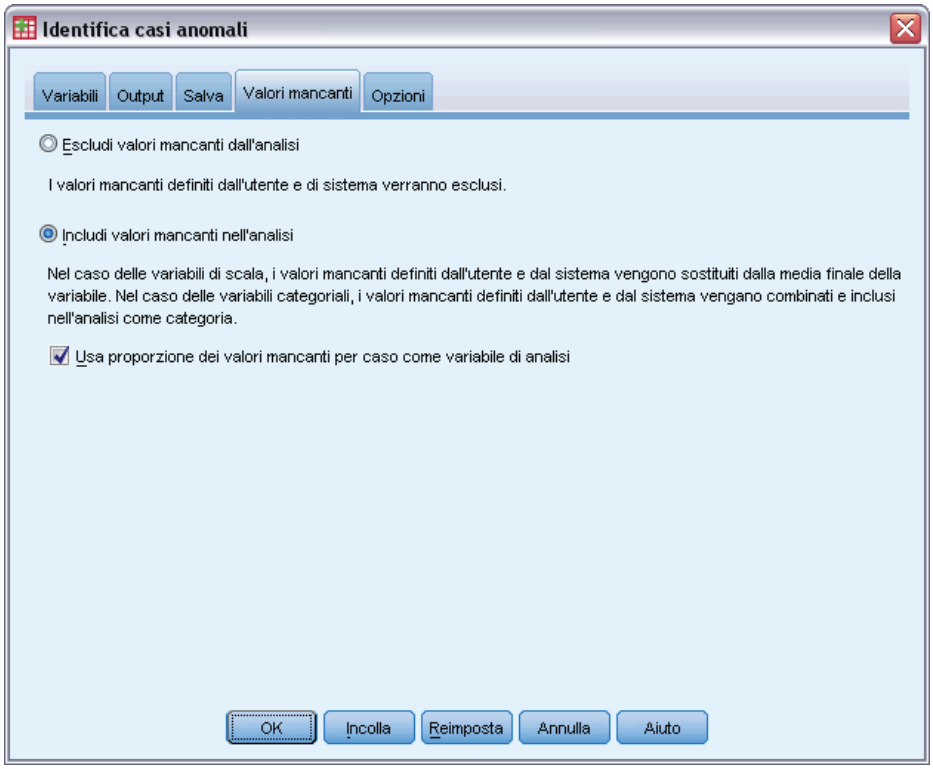


- Selezionare Indice di anomalia, Gruppi equivalenti e Motivi.

  Il salvataggio di questi risultati permette di creare un grafico a dispersione che riepiloga i risultati.

- Fare clic sulla scheda Valori mancanti.

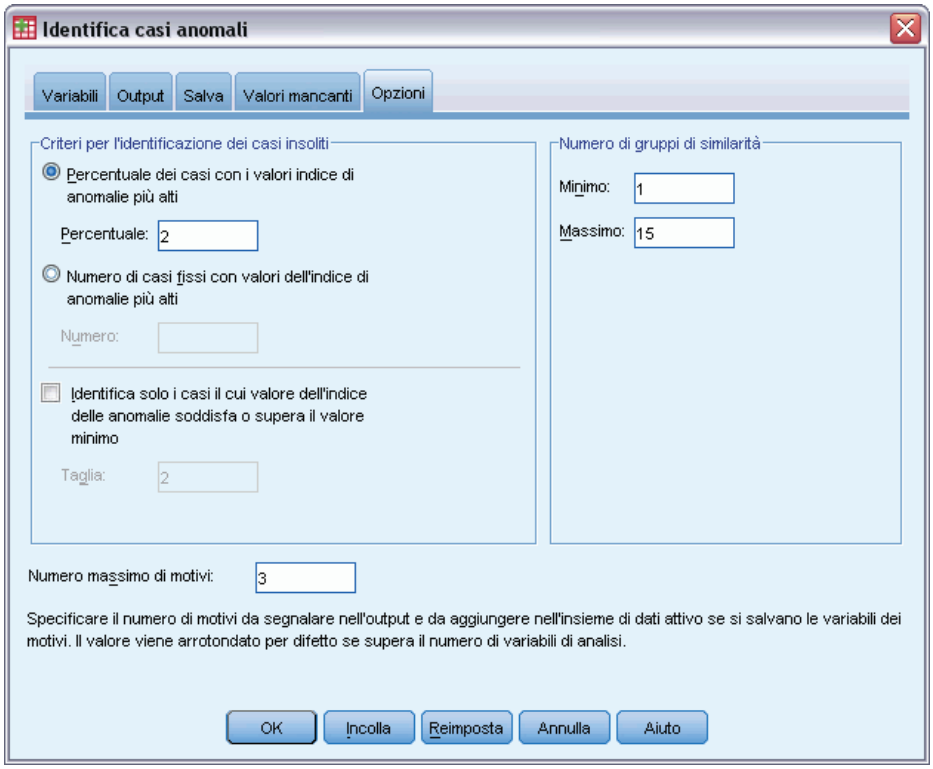
Figura 9-4
Scheda Valori mancanti della finestra di dialogo Identifica casi anomali



- Selezionare Includi valori mancanti nell'analisi. Questo processo è necessario a causa dell'elevato numero di valori mancanti definiti dall'utente necessario per gestire i pazienti che sono deceduti prima o durante il trattamento. All'analisi viene aggiunta una variabile aggiuntiva che misura la proporzione di casi mancanti per casi come variabile di scala.

- Fare clic sulla scheda Opzioni.

Figura 9-5
Scheda Opzioni della finestra di dialogo Identifica casi anomali



- Immettere 2 come percentuale di casi da considerare anomali.
- Deselezionare Identifica solo i casi con valori indice di anomalia uguale o maggiore del valore minimo.
- Digitare 3 come numero massimo dei motivi.
- Fare clic su OK.

Riepilogo dei casi

Figura 9-6
Riepilogo dell'elaborazione dei casi

		N	% di casi combinati	% del totale
ID similarità	1	710	67,7%	67,7%
	2	90	8,6%	8,6%
	3	248	23,7%	23,7%
Combinata		1048	100,0%	100,0%
Totale		1048		100,0%

Ciascun caso viene classificato in un gruppo equivalente di casi simili. Il riassunto dell'elaborazione casi mostra il numero di gruppi equivalenti creati nonché il numero e la percentuale di casi in ciascun gruppo equivalente.

Elenco Indice dei casi anomali

Figura 9-7
Elenco Indice dei casi anomali

Caso	patid	Indice delle anomalie
843	7840326167	2,837
510	0714726620	2,022
623	6553808330	2,014
501	6461046805	2,002
607	1077125669	1,897
884	2260043998	1,889
614	4030164769	1,869
241	1038840465	1,865
13	2191527525	1,826
172	4458028382	1,786
705	1336411777	1,778
651	4103977868	1,767
384	2247641363	1,767
839	0437454972	1,766
861	9746101913	1,757
19	7237535360	1,756
806	4391632997	1,756
871	6961938294	1,739
239	7315965190	1,738
887	6044244232	1,737
245	0816869249	1,736

L'indice delle anomalie è una misura che riflette le anomalie di un caso rispetto al gruppo di appartenenza. Viene visualizzato il 2% dei casi dell'indice di anomalia con i valori più alti, unitamente ai numeri e agli ID dei casi. Vengono elencati ventuno casi, con valori compresi tra 1,736 e 2,837. Poiché la differenza del valore dell'indice di anomalia tra i primi e i secondi casi dell'elenco è molto alta, è probabile che il caso 843 sia anomalo. Gli altri casi devono essere esaminati caso per caso.

Elenco ID casi anomali equivalenti

Figura 9-8
Elenco ID casi anomali equivalenti

Caso	patid	ID similarità	Dimensione del gruppo di similarità	Dimensione percentuale del gruppo di similarità
843	7840326167	3	248	23,7%
510	0714726620	3	248	23,7%
623	6553808330	3	248	23,7%
501	6461046805	3	248	23,7%
607	1077125669	3	248	23,7%
884	2260043998	3	248	23,7%
614	4030164769	3	248	23,7%
241	1038840465	3	248	23,7%
13	2191527525	3	248	23,7%
172	4458028382	3	248	23,7%
705	1336411777	1	710	67,7%
651	4103977868	1	710	67,7%
384	2247641363	3	248	23,7%
839	0437454972	3	248	23,7%
861	9746101913	3	248	23,7%
19	7237535360	1	710	67,7%
806	4391632997	1	710	67,7%
871	6961938294	1	710	67,7%
239	7315965190	3	248	23,7%
887	6044244232	1	710	67,7%
245	0816869249	3	248	23,7%

I casi potenzialmente anomali vengono visualizzati insieme alle informazioni del gruppo equivalente di appartenenza. I primi 10 casi e complessivamente 15 casi appartengono al gruppo equivalente 3, mentre i restanti appartengono al gruppo equivalente 1.

Elenco Motivi anomalie

Figura 9-9
Elenco Motivi anomalie

Motivo: 1

Caso	patid	Variabile motivo	Impatto delle variabili	Valore delle variabili	Valore normale
843	7840326167	cost	,411	200,51	19,83
510	0714726620	cost	,120	96,59	19,83
623	6553808330	cost	,175	114,01	19,83
501	6461046805	barthel1	,084	80	(Valore mancante)
607	1077125669	cost	,126	96,11	19,83
884	2260043998	cost	,138	99,73	19,83
614	4030164769	rankin1	,085	3	(Valore mancante)
241	1038840465	barthel1	,115	25	(Valore mancante)
13	2191527525	barthel1	,118	40	(Valore mancante)
172	4458028382	barthel1	,120	100	(Valore mancante)
705	1336411777	cost	,244	198,25	42,47
651	4103977868	barthel1	,064	30	95
384	2247641363	barthel1	,122	20	(Valore mancante)
839	0437454972	barthel1	,109	95	(Valore mancante)
861	9746101913	barthel1	,102	70	(Valore mancante)
19	7237535360	barthel3	,080	5	100
806	4391632997	barthel2	,088	10	100
871	6961938294	barthel1	,094	5	95
239	7315965190	rankin1	,092	3	(Valore mancante)
887	6044244232	stroke1	,066	1	0
245	0816869249	barthel1	,124	5	(Valore mancante)

Le variabili del motivo sono variabili che più contribuiscono alla classificazione di un caso come anomalo. Viene visualizzata la variabile del motivo principale per ciascun caso anomalo, unitamente al suo impatto, al valore del caso e alla norma del gruppo equivalente. La norma del gruppo equivalente *(Valore mancante)* di una variabile categoriale indica che più casi nel gruppo equivalente contenevano un valore mancante per la variabile.

La statistica di impatto della variabile rappresenta il contributo proporzionale della variabile del motivo alla deviazione del caso rispetto al gruppo equivalente. Se l'analisi contiene 38 variabili, compresa la variabile mancante della proporzione, l'impatto atteso della variabile è pari a 1/38 = 0,026. L'impatto della variabile *costo* sul caso 843 è 0,411, ovvero un valore relativamente alto. Il valore della variabile *costo* per il caso 843 è 200,51 contro la media di 19,83 dei casi del gruppo equivalente 3.

Le selezioni della finestra di dialogo indicavano che erano richiesti i risultati per i primi tre motivi.

- Per visualizzare i risultati di altri motivi, fare doppio clic sulla tabella per abilitarla.
- Spostare *Motivo* dalla dimensione di strato a quella di riga.

Figura 9-10
Elenco Motivi anomalie (primi 8 casi)

Caso	Motivo	patid	Variabili motivo	Impatto delle variabili	Valore delle variabili	Valore normale
843	1	7840326167	cost	.411	200.51	19.83
	2	7840326167	barthel1	.076	65	(Valore mancante)
	3	7840326167	rankin1	.044	2	(Valore mancante)
510	1	0714726620	cost	.120	96.59	19.83
	2	0714726620	barthel1	.083	80	(Valore mancante)
	3	0714726620	rehab	.068	3	(Valore mancante)
623	1	6553808330	cost	.175	114.01	19.83
	2	6553808330	surgery	.089	2	(Valore mancante)
	3	6553808330	barthel1	.089	70	(Valore mancante)
501	1	6461046805	barthel1	.084	80	(Valore mancante)
	2	6461046805	rehab	.068	3	(Valore mancante)
	3	6461046805	rankin1	.063	1	(Valore mancante)
607	1	1077125669	cost	.126	96.11	19.83
	2	1077125669	barthel1	.094	85	(Valore mancante)
	3	1077125669	rehab	.072	3	(Valore mancante)
884	1	2260043998	cost	.138	99.73	19.83
	2	2260043998	barthel1	.114	65	(Valore mancante)
	3	2260043998	rehab	.072	3	(Valore mancante)
614	1	4030164769	barthel1	.085	45	(Valore mancante)
	2	4030164769	rankin1	.085	3	(Valore mancante)
	3	4030164769	recbart1	.062	2	(Valore mancante)

Questa configurazione permette di confrontare facilmente i contributi dei primi tre motivi per ciascun caso. Il caso 843, come si sospettava, è considerato anomalo a causa del valore elevato della variabile *costo*. Per contro nessuno motivo singolo contribuisce più dello 0,10 a rendere il caso 501 anomalo.

Norme delle variabili di scala

Figura 9-11
Norme delle variabili di scala

		ID similarità			
		1	2	3	Combinata
los_rehab	Media	16,55	16,39	15,91	16,39
	Deviazione standard	12,596	,000	6,834	10,887
cost	Media	42,4673	3,5089	19,8273	33,7641
	Deviazione standard	26,45401	,50997	20,17309	27,31266
Proporzione mancante	Media	,006	,541	,354	,134
	Deviazione standard	,021	,000	,083	,197

Le norme delle variabili di scala indicano la deviazione media e standard di ciascuna variabile, sia a livello generale che a livello di gruppi equivalenti. Il confronto dei valori permette di identificare le variabili che contribuiscono alla formazione del gruppo equivalente.

Ad esempio la media di *Durata della riabilitazione* è abbastanza costante in tutti e tre i gruppi equivalenti, quindi non contribuisce alla formazione dei gruppi equivalenti. *Costi totali di cura e riabilitazione in migliaia* e *Proporzione mancate* forniscono invece indicazioni utili sull'appartenenza al gruppo equivalente. Il gruppo equivalente 1 ha il costo medio più alto e il numero minore di valori mancanti. Il gruppo equivalente 2 ha costi molto bassi e molti valori mancanti. Il gruppo equivalente 3 ha costi e valori mancanti intermedi.

Questa organizzazione indica che il gruppo equivalente 2 è costituito da pazienti che risultavano deceduti all'arrivo e che hanno richiesto costi minimi e contribuito a classificare le variabili relative alla cura e alla riabilitazione come mancanti. Il gruppo equivalente 3 contiene molti pazienti che sono deceduti durante il trattamento e che quindi hanno inciso sui costi di cura ma non su quelli di riabilitazione, nonché contribuito a rendere le variabili relative alla riabilitazione mancanti. Il gruppo equivalente 1 comprende probabilmente pazienti che sono sopravvissuti alle cure e alla riabilitazione e che hanno inciso significativamente sui costi.

Norme delle variabili categoriali

Figura 9-12
Norme delle variabili categoriali (prime 10 variabili)

		ID similarità			
		1	2	3	Combinata
agecat	Categoria più popolare	2	3	2	2
	Frequenza	277	25	81	383
	Percentuale	39,0%	27,8%	32,7%	36,5%
gender	Categoria più popolare	0	0	1	0
	Frequenza	361	46	126	529
	Percentuale	50,8%	51,1%	50,8%	50,5%
active	Categoria più popolare	1	0	0	0
	Frequenza	373	55	139	531
	Percentuale	52,5%	61,1%	56,0%	50,7%
obesity	Categoria più popolare	0	0	0	0
	Frequenza	555	67	178	800
	Percentuale	78,2%	74,4%	71,8%	76,3%
diabetes	Categoria più popolare	0	0	0	0
	Frequenza	665	80	219	964
	Percentuale	93,7%	88,9%	88,3%	92,0%
bp	Categoria più popolare	1	1	1	1
	Frequenza	445	49	139	633
	Percentuale	62,7%	54,4%	56,0%	60,4%
af	Categoria più popolare	0	0	0	0
	Frequenza	641	83	216	940
	Percentuale	90,3%	92,2%	87,1%	89,7%
smoker	Categoria più popolare	0	0	0	0
	Frequenza	578	69	179	826
	Percentuale	81,4%	76,7%	72,2%	78,8%
choles	Categoria più popolare	0	0	0	0
	Frequenza	406	52	136	594
	Percentuale	57,2%	57,8%	54,8%	56,7%
angina	Categoria più popolare	0	0	0	0
	Frequenza	493	52	167	712
	Percentuale	69,4%	57,8%	67,3%	67,9%

Le norme delle variabili categoriali sono simili alle norme delle variabili di scala, ma a differenza di quest'ultime forniscono indicazioni sulla categoria modale (più utilizzata) e sul numero e sulla percentuale di casi del gruppo equivalente che rientra nella categoria specificata. In questo caso il confronto dei valori può essere più complesso. Da un'analisi sommaria, può sembrare che *Sesso* contribuisca maggiormente alla formazione di cluster rispetto a *Fumatore* perché la categoria modale di *Fumatore* è la stessa per tutti e tre i gruppi equivalenti, mentre la categoria modale di *Sesso* differisce nel gruppo equivalente 3. Tuttavia, poiché *Sesso* ha solo due valori, è

possibile supporre che il 49,2% dei casi nel gruppo equivalente 3 abbia un valore 0, ovvero un valore simile alle percentuali degli altri gruppi equivalenti. Le percentuali relative a *Fumatore* sono invece comprese tra il 72,2% e l'81,4%.

Figura 9-13
Norme delle variabili categoriali (variabili selezionate)

		ID similarità			
		1	2	3	Combinata
doa	Categoria più popolare	0	1	0	0
	Frequenza	710	90	248	958
	Percentuale	100,0%	100,0%	100,0%	91,4%
rankin0	Categoria più popolare	0	(Valore mancante)	5	5
	Frequenza	166	90	104	193
	Percentuale	23,4%	100,0%	41,9%	18,4%
catscan	Categoria più popolare	0	(Valore mancante)	0	0
	Frequenza	607	90	184	791
	Percentuale	85,5%	100,0%	74,2%	75,5%
clotsolv	Categoria più popolare	2	(Valore mancante)	0	2
	Frequenza	318	90	129	394
	Percentuale	44,8%	100,0%	52,0%	37,6%
dhosp	Categoria più popolare	0	(Valore mancante)	1	0
	Frequenza	710	90	171	787
	Percentuale	100,0%	100,0%	69,0%	75,1%
result	Categoria più popolare	1	(Valore mancante)	1	1
	Frequenza	524	90	96	620
	Percentuale	73,8%	100,0%	38,7%	59,2%
surgery	Categoria più popolare	0	(Valore mancante)	(Valore mancante)	0
	Frequenza	323	90	171	369
	Percentuale	45,5%	100,0%	69,0%	35,2%
rehab	Categoria più popolare	0	(Valore mancante)	(Valore mancante)	0
	Frequenza	278	90	171	314
	Percentuale	39,2%	100,0%	69,0%	30,0%

I sospetti messi in evidenza dalle norme delle variabili di scala sono confermati dai risultati della tabella delle norme delle variabili categoriali. Il gruppo equivalente 2 è costituito interamente da pazienti che risultavano deceduti al momento dell'arrivo e in cui tutte le variabili relative alla cura e alla riabilitazione risultano mancanti. La maggior parte dei pazienti del gruppo equivalente 3 (69,0%) è deceduta durante la cura, quindi la categoria modale per la variabile relativa alla riabilitazione è *(Valore mancante)*.

Riassunto Indice delle anomalie

Figura 9-14
Riassunto Indice delle anomalie

	N nell'elenco anomalie	Minimo	Massimo	Media	Deviazione standard
Indice delle anomalie	21	1,736	2,837	1,872	,240

N nell'elenco Anomalie è determinato dalla specifica: la percentuale delle anomalie è 2%

La tabella fornisce dati statistici riassuntivi dei valori dell'indice delle anomalie per i casi contenuti nell'elenco delle anomalie.

Riassunto Motivi

Figura 9-15
Riassunto Motivi (variabili relative alla cura e alla riabilitazione)

	Occorenza come motivo		Statistische dell'impatto delle variabili			
	Frequenza	Percent.	Minimo	Massimo	Media	Deviazione std.
doa	0	,0%	.	.	.	.
rankin0	0	,0%	.	.	.	.
catscan	0	,0%	.	.	.	.
clotsolv	0	,0%	.	.	.	.
dhosp	0	,0%	.	.	.	.
result	0	,0%	.	.	.	.
surgery	0	,0%	.	.	.	.
rehab	0	,0%	.	.	.	.
rankin1	0	,0%	.	.	.	.
rankin2	0	,0%	.	.	.	.
rankin3	0	,0%	.	.	.	.
barthel1	13	61,9%	,064	,124	,100	.021
barthel2	1	4,8%	,088	,088	,088	.
barthel3	1	4,8%	,080	,080	,080	.
recbart1	0	,0%	.	.	.	.
recbart2	0	,0%	.	.	.	.
recbart3	0	,0%	.	.	.	.
stroke1	0	,0%	.	.	.	.
stroke2	0	,0%	.	.	.	.
stroke3	0	,0%	.	.	.	.
los_rehab	0	,0%	.	.	.	.
cost	6	28,6%	,120	,411	,202	.112
Proporzione mancante	0	,0%	.	.	.	.
Globale	21	100,0%	,064	,411	,127	,076

La tabella riepiloga, per ciascuna variabile dell'analisi, il ruolo della variabile come motivo principale. La maggior parte delle variabili comprese tra *Deceduto all'arrivo* e *Riabilitazione post-evento* non sono il motivo principale per cui i casi sono stati inclusi nell'elenco delle anomalie. *Indice di Barthel dopo 1 mese* è il motivo più frequente, seguito da *Costi totali di cura e riabilitazione in migliaia*. Vengono riassunte le statistiche di impatto delle variabili, unitamente

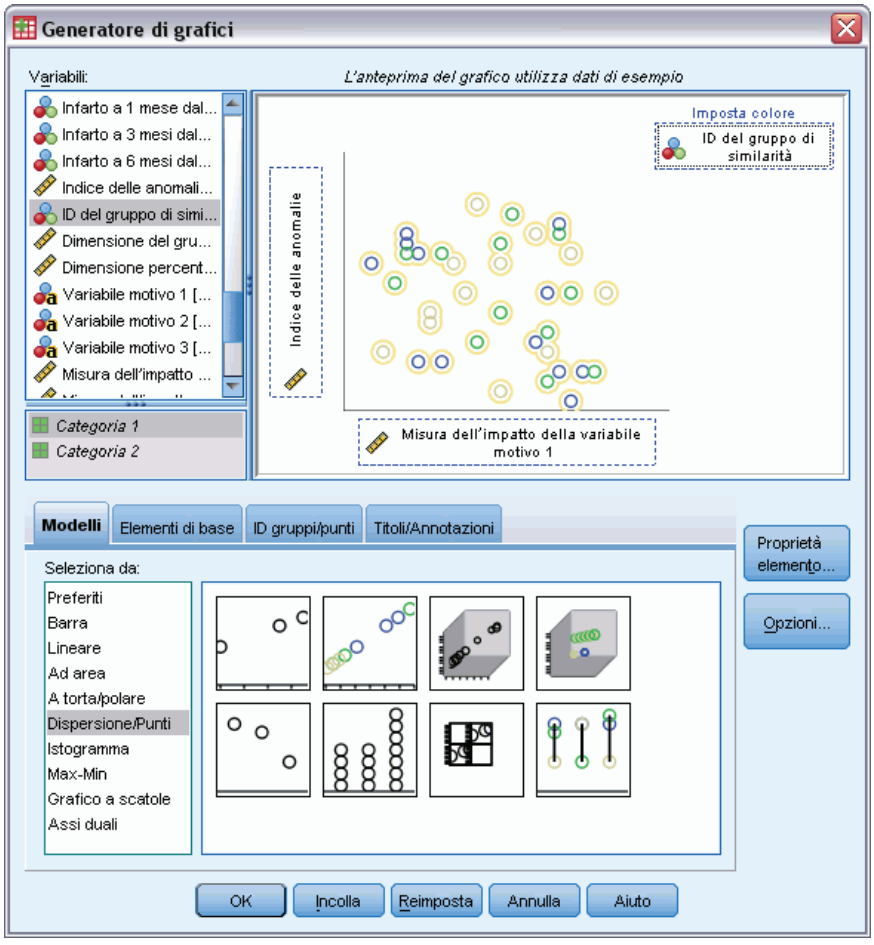
all'impatto minimo, massimo e medio di ciascuna variabile e alla deviazione standard delle variabili che costituiscono un motivo per più casi.

Grafico a dispersione dell'indice delle anomalie per impatto della variabile

Benché le tabelle contengano molte informazioni utili, può essere difficile coglierne le relazioni. Le variabili salvate possono essere utilizzate per creare un grafico che semplifichi il processo.

- Per creare il grafico a dispersione, selezionare dai menu:
  Grafici > Generatore di grafici...

Figura 9-16
Finestra di dialogo Generatore di grafici

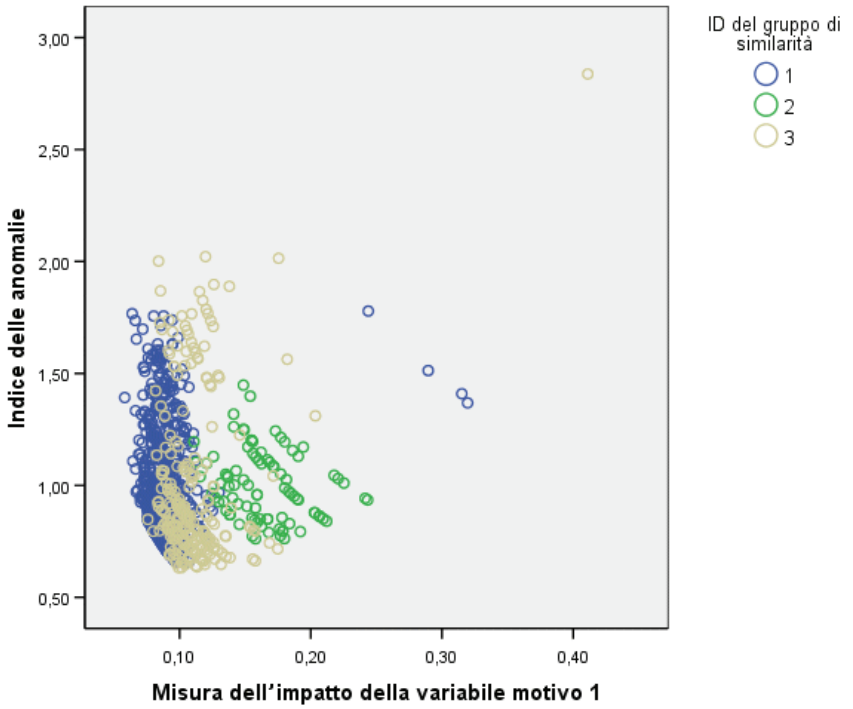


- Selezionare il modello Dispersione/Punti e trascinare l'icona A dispersione raggruppato nel disegno.
- Selezionare *Indice di anomalia* come variabile *y* e *Misura dell'impatto della variabile motivo 1* come variabile *x*.
- Selezionare *ID gruppo equivalente* come variabile da utilizzare per l'impostazione dei colori.

▶ Fare clic su OK.

Queste selezioni permettono di creare il grafico a dispersione.

Figura 9-17
Grafico a dispersione dell'indice di anomalia per la misura dell'impatto della prima variabile del motivo



L'analisi del grafico permette di giungere a varie conclusioni:

- il caso nell'angolo in alto a destra appartiene al gruppo equivalente 3 ed è sia quello più anomalo che quello con il contributo più alto fornito da una singola variabile.
- Se ci si sposta lungo l'asse *y*, si nota che ci sono tre casi del gruppo equivalente 3 i cui valori dell'indice di anomalia sono leggermente superiori a 2,00. Questi casi devono essere esaminati con maggiore attenzione perché potrebbero essere anomali.
- Se ci si sposta lungo l'asse *x*, si nota che ci sono quattro casi del gruppo equivalente 1 in cui l'impatto delle variabili è approssimativamente compreso nell'intervallo tra 0,23 e 0,33. Questi casi devono essere esaminati con maggiore attenzione poiché questi valori separano i casi dal corpo principale dei punti nel grafico.
- Il gruppo equivalente 2 appare abbastanza omogeneo nel senso che i valori dell'indice di anomalia e dell'impatto delle variabili non si discosta molto dalle tendenze del corpo centrale.

Riepilogo

La procedura Identifica casi anomali ha permesso di individuare molti casi che richiedono ulteriori analisi. Questi sono casi che non sarebbe stato possibile identificare con altre procedure di convalida, poiché i rapporti tra le variabili (oltre ai valori delle variabili stesse) generano casi anomali.

L'elemento sconcertante è che i gruppi equivalenti sono principalmente costituiti da due variabili: *Deceduto all'arrivo* e *Deceduto in ospedale*. Con ulteriori analisi sarebbe possibile esaminare l'effetto della creazione forzata di un maggior numero di gruppi equivalenti oppure eseguire uno studio specifico sui pazienti che sono sopravvissuti al trattamento.

Procedure correlate

La procedura Identifica casi anomali è uno strumento utile per rilevare casi anomali nei file di dati.

- La procedura Convalida i dati permette di identificare casi, variabili e valori dati sospetti e non validi nel file di dati attivo.

Capitolo
10

Categorizzazione ottimale

La procedura Categorizzazione ottimale discretizza una o più variabili di scala (denominate **variabili di input di categorizzazione**) distribuendo i valori di ogni variabile in intervalli. La formazione di intervalli risulta migliore rispetto a una variabile guida categoriale che esegue la "supervisione" del processo di categorizzazione, in quanto gli intervalli possono essere utilizzati al posto dei valori originali dei dati per ulteriori analisi in procedure che richiedono o preferiscono variabili categoriali.

Algoritmo di categorizzazione ottimale

I passaggi fondamentali dell'algoritmo di categorizzazione ottimale sono caratterizzati come segue:

Preprocesso (facoltativo). La variabile di input di categorizzazione è divisa in *n* intervalli (dove *n* è specificato dall'utente) e ogni intervallo contiene lo stesso numero di casi oppure il numero di casi il più possibile simile.

Identificazione dei potenziali punti di divisione. Ogni valore diverso dell'input di categorizzazione che non appartiene alla stessa categoria della variabile guida come il valore diverso maggiore successivo della variabile di input di categorizzazione è un potenziale punto di divisione.

Selezione dei punti di divisione. Il potenziale punto di divisione che produce il maggior guadagno di informazioni viene valutato mediante il criterio di accettazione MDLP. Ripetere fino all'esaurimento dei potenziali punti di divisione accettati. I punti di divisione accettati definiscono i punti finali degli intervalli.

Utilizzo della categorizzazione ottimale per la discretizzazione dei dati dei mutuatari

Nell'ambito del tentativo di una banca di ridurre il tasso di inadempienza nel rimborso di un prestito, un funzionario mutui ha raccolto informazioni finanziarie e demografiche su clienti passati e presenti con l'intenzione di creare un modello per la previsione della probabilità di inadempienza. Numerosi predittori potenziali sono variabili di scala ma il funzionario desidera poter considerare modelli che danno buoni risultati con predittori categoriali.

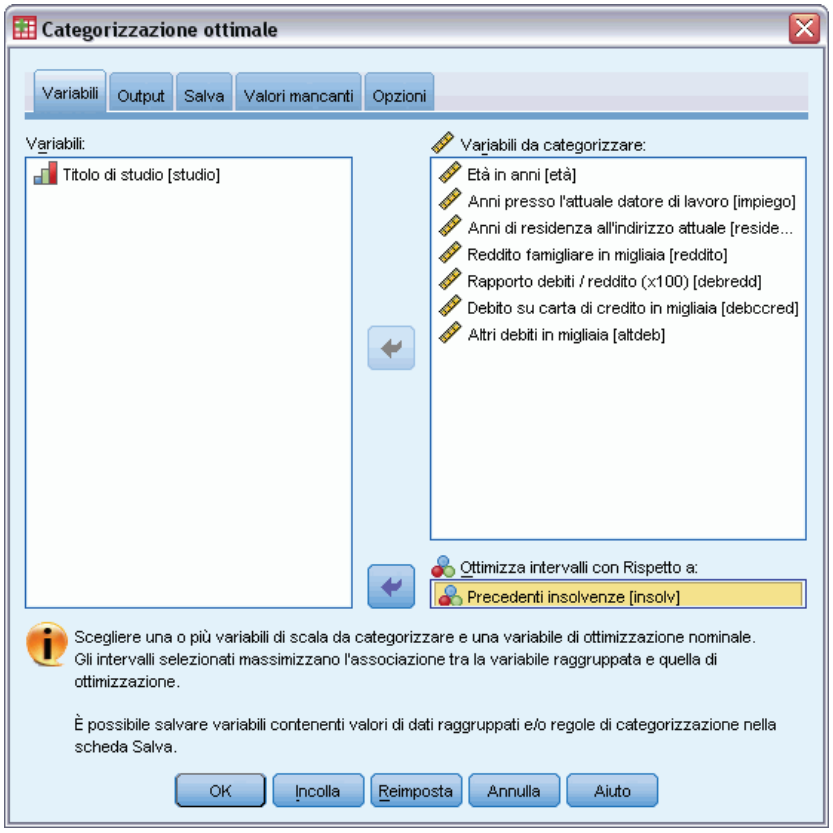
I dati relativi a 5000 vecchi clienti sono contenuti nel file *bankloan_binning.sav*. Per ulteriori informazioni, vedere l'argomento File di esempio in l'appendice A a pag. 137.Utilizzare la categorizzazione ottimale per generare regole di categorizzazione per i predittori di scala, quindi

utilizzare tali regole per elaborare *bankloan.sav*. L'insieme di dati sottoposto a preprocesso può quindi essere utilizzato per creare un modello predittivo.

Esecuzione dell'analisi

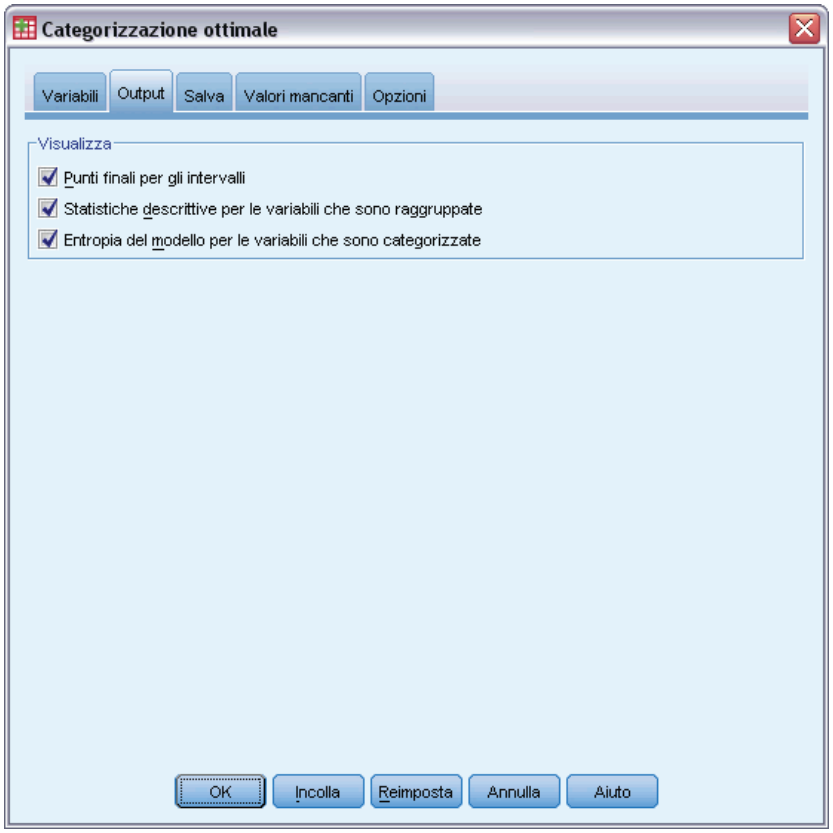
- Per eseguire l'analisi Categorizzazione ottimale, dai menu scegliere:
  Trasforma > Categorizzazione ottimale...

Figura 10-1
Scheda Variabili della finestra di dialogo Categorizzazione ottimale



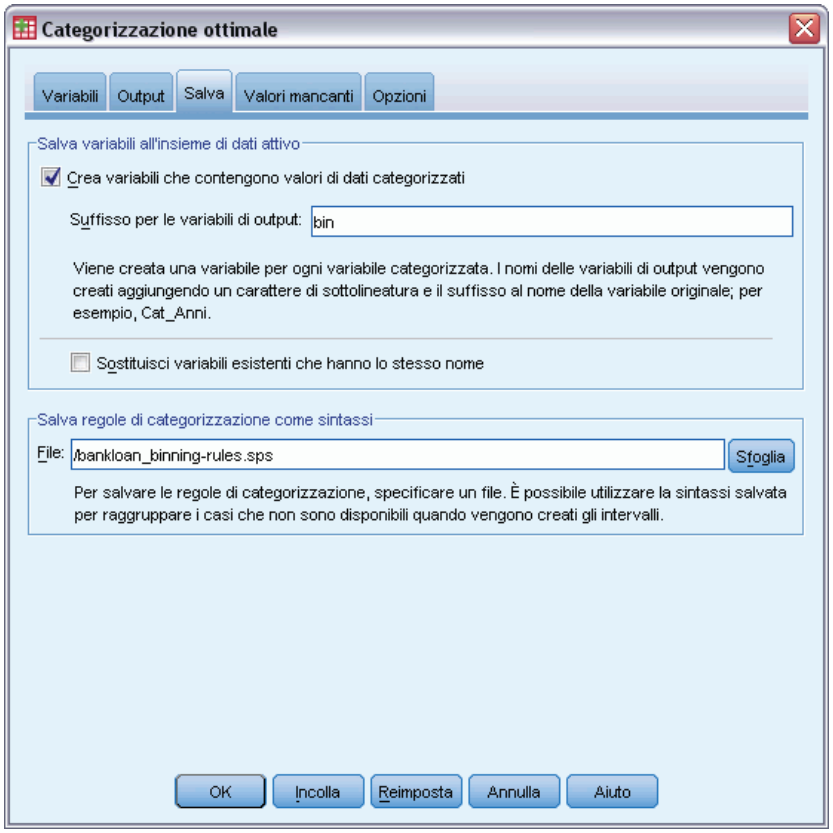
- Selezionare *Età in anni* e *Anni di permanenza nell'impiego attuale* fino a *Altri indebitamenti in migliaia* come variabili da categorizzare.
- Selezionare *Già inadempiente* come variabile guida.
- Fare clic sulla scheda Output.

Figura 10-2
Scheda Output della finestra di dialogo Categorizzazione ottimale



- Selezionare Statistiche descrittive e Entropia del modello per le variabili che sono categorizzate.
- Fare clic sulla scheda Salva.

Figura 10-3
Scheda Salva della finestra di dialogo Categorizzazione ottimale



- Selezionare Crea variabili che contengono valori di dati categorizzati.
- Immettere un percorso e un nome per il file di sintassi che conterrà le regole di categorizzazione generate. Nell'esempio è stato usato il file */bankloan_binning-rules.sps*.
- Fare clic su OK.

Tali selezioni generano la sintassi seguente:

```
* Categorizzazione ottimale.
OPTIMAL BINNING
  /VARIABLES GUIDE=default BIN=age employ address income debtinc creddebt
  othdebt SAVE=YES (INTO=age_bin employ_bin address_bin income_bin debtinc_bin
  creddebt_bin othdebt_bin)
  /CRITERIA METHOD=MDLP
   PREPROCESS=EQUALFREQ (BINS=1000)
   FORCEMERGE=0
   LOWERLIMIT=INCLUSIVE
   LOWEREND=UNBOUNDED
   UPPEREND=UNBOUNDED
  /MISSING SCOPE=PAIRWISE
  /OUTFILE RULES='/bankloan_binning-rules.sps'
  /PRINT ENDPOINTS DESCRIPTIVES ENTROPY.
```

- La procedura discretizza le variabili di input di categorizzazione *età*, *impiego*, *indirizzo*, *reddito*, *debred*, *creddeb* e *altrideb* utilizzando la catogorizzazione MDLP con la variabile guida *default*.
- I valori discretizzati per queste variabili verranno archiviati nelle nuove variabili *age_bin*, *employ_bin*, *address_bin*, *income_bin*, *debtinc_bin*, *creddebt_bin* e *othdebt_bin*.
- Se una variabile di input di categorizzazione presenta oltre 1000 valori distinti, il metodo delle frequenze uguali ridurrà il numero a 1000 prima di eseguire la categorizzazione MDLP.
- La sintassi del comando che rappresenta le regole di categorizzazione viene salvata nel file *\bankloan_binning-rules.sps*.
- Punti finali intervallo, statistiche descrittive e valori di entropia del modello sono necessari per le variabili di input di categorizzazione.
- Altri criteri di categorizzazione vengono impostati sui valori predefiniti.

Statistiche descrittive

Figura 10-4
Statistiche descrittive

	N	Minimo	Massimo	Numero di valori distinti	Numero di raggruppamenti
Age in years	5000	20	58	39	2
Years with current employer	5000	0	38	39	4
Years at current address	5000	0	37	38	3
Household income in thousands	5000	12,10	2461,70	1100	2
Debt to income ratio (x100)	5000	,08	44,62	2060	5
Credit card debt in thousands	5000	,01	139,58	5000	4
Other debt in thousands	5000	,01	416,52	4999	2

Nella tabella delle statistiche descrittive sono contenute informazioni riepilogative sulle variabili di input di categorizzazione. Le prime quattro colonne riguardano i valori prima della categorizzazione.

- N è il numero di casi utilizzati nell'analisi. Quando viene utilizzata l'eliminazione listwise dei valori mancanti, tale valore deve essere costante tra le varie variabili. Quando viene utilizzata la gestione pairwise dei dati mancanti non è necessario che questo valore sia costante. Poiché questo insieme di dati non presenta valori mancanti, il valore coincide semplicemente con il numero di casi.
- Nelle colonne Minimum e Maximum sono indicati i valori minimi e massimi (prima della categorizzazione) dell'insieme di dati per ogni variabile di input di categorizzazione. Oltre a dare un'idea dell'intervallo osservato di valori per ogni variabile, essi possono risultare utili per rilevare valori al di fuori dell'intervallo previsto.
- I valori nella colonna Number of Distinct Values (numero di valori diversi) indicano le variabili sottoposte a preprocesso mediante l'algoritmo delle frequenze uguali. Per impostazione predefinita le variabili con oltre 1000 valori diversi (da *Reddito familiare in migliaia* a *Altri*

indebitamenti in migliaia) vengono raggruppate in 1000 intervalli diversi. Gli intervalli sottoposti a preprocesso vengono quindi sottoposti a categorizzazione rispetto alla variabile guida mediante MDLP. È possibile controllare la funzione di preprocesso nella scheda Opzioni.

- Il numero di intervalli è il numero finale di categorizzazioni generate dalla procedura ed è di molto inferiore al numero di valori diversi.

Entropia modello

Figura 10-5
Entropia del modello

	Entropia del modello
Age in years	,788
Years with current employer	,754
Years at current address	,781
Household income in thousands	,803
Debt to income ratio (x100)	,711
Credit card debt in thousands	,776
Other debt in thousands	,801

Entropia del modello minore indica un'accuratezza predittiva maggiore delle variabili raggruppate sulla variabile guida Previously defaulted.

L'entropia del modello consente di valutare l'utilità di ogni variabile in un modello predittivo relativamente alla probabilità di inadempienza.

- Il miglior predittore possibile è quello che, per ogni intervallo generato, contiene casi con lo stesso valore della variabile guida, di conseguenza la variabile guida può essere prevista perfettamente. Tale predittore presenta un'entropia di modello indefinita. Solitamente ciò non si verifica nelle situazioni reali e potrebbe indicare l'esistenza di problemi a livello di qualità dei dati.
- Il peggior predittore possibile è quello che si limita a ipotizzare. Il valore della relativa entropia di modello dipende dai dati. In questo insieme di dati 1256 (o 0,2512) dei 5000 clienti totali sono risultati inadempienti e 3744 (o 0,7488) hanno pagato puntualmente. Pertanto il peggior predittore possibile avrebbe l'entropia di $-0{,}2512 \times \log_2(0{,}2512) - 0{,}7488 \times \log_2(0{,}7488) = 0{,}8132$.

È difficile fare a meno di concludere che le variabili con valori di entropia di modello più bassi sono i predittori migliori, dal momento che ciò che costituisce un buon valore di entropia di modello dipende dall'applicazione e dai dati. In questo caso sembra che le variabili con un numero maggiore di intervalli generati, rispetto al numero di categorie diverse, abbiano valori di entropia di modello inferiori. Un'ulteriore valutazione di queste variabili di input di categorizzazione come predittori deve essere eseguita utilizzando modelli predittivi, che dispongono di strumenti più adeguati per la selezione di variabili.

Riepiloghi di categorizzazione

Nel riepilogo di categorizzazione sono contenuti i limiti degli intervalli generati e il conteggio delle frequenze di ogni intervallo in base a valori della variabile guida. Viene prodotta una tabella riepilogativa di categorizzazione separata per ogni variabile di input di categorizzazione.

Figura 10-6
Riepilogo di categorizzazione per Età in anni

	Punto finale		Numero di casi per livello di Previously defaulted		
Bin	Inferiore	Superiore	No	Yes	Totale
1	[a]	32	1129	639	1768
2	32	[a]	2615	617	3232
Totale			3744	1256	5000

Ciascun bin viene calcolato come Inferiore <= Age in years < Superiore.

a. Illimitato

Il riepilogo per *Età in anni* indica che 1768 clienti, tutti di età inferiore a 32 anni, vengono collocati in Intervallo 1 mentre i restanti 3232, tutti di età superiore a 32 anni, vengono collocati in Intervallo 2. La proporzione dei clienti che in precedenza si è dimostrata inadempiente è molto maggiore in Categorizzazione 1 (639/1768=0,361) rispetto a Categorizzazione 2 (617/3232=0,191).

Figura 10-7
Riepilogo di categorizzazione per Reddito familiare in migliaia

	Punto finale		Numero di casi per livello di Previously defaulted		
Bin	Inferiore	Superiore	No	Yes	Totale
1	[a]	26,70	1054	513	1567
2	26,70	[a]	2690	743	3433
Totale			3744	1256	5000

Ciascun bin viene calcolato come Inferiore <= Household income in thousands < Superiore.

a. Illimitato

Il riepilogo per *Reddito familiare in migliaia* mostra uno schema simile, con un unico punto di divisione in corrispondenza di 26,70 e una proporzione maggiore di clienti precedentemente inadempienti in Intervallo 1 (513/1567=0,327) rispetto a Intervallo 2 (743/3433=0,216). Come prevedibile in base alle statistiche dell'entropia di modello, la differenza in queste proporzioni non è grande come per *Età in anni*.

Figura 10-8
Riepilogo di categorizzazione per Altri indebitamenti in migliaia

	Punto finale		Numero di casi per livello di Previously defaulted		
Bin	Inferiore	Superiore	No	Yes	Totale
1	[a]	2,19	2161	539	2700
2	2,19	[a]	1583	717	2300
Totale			3744	1256	5000

Ciascun bin viene calcolato come Inferiore <= Other debt in thousands < Superiore.

a. Illimitato

Il riepilogo per *Altri indebitamenti in migliaia* mostra uno schema simile, con un unico punto di divisione in corrispondenza di 2,19 e una proporzione minore di clienti precedentemente inadempienti in Intervallo 1 (539/2700=0,200) rispetto a Intervallo 2 (717/2300=0,312). Di nuovo, come prevedibile in base alle statistiche dell'entropia di modello, la differenza in queste proporzioni non è grande come per *Età in anni*.

Figura 10-9
Riepilogo di categorizzazione per Anni di permanenza nell'impiego attuale

	Punto finale		Numero di casi per livello di Previously defaulted		
Bin	Inferiore	Superiore	No	Yes	Totale
1	[a]	3	629	478	1107
2	3	8	1066	461	1527
3	8	18	1471	268	1739
4	18	[a]	578	49	627
Totale			3744	1256	5000

Ciascun bin viene calcolato come Inferiore <= Years with current employer < Superiore.

a. Illimitato

Il riepilogo per *Anni di permanenza nell'impiego attuale* mostra uno schema di proporzioni decrescenti di soggetti inadempienti man mano che aumentano i numeri di intervalli.

Intervallo	**Proporzione di inadempienti**
1	0.432
2	0.302
3	0.154
4	0.078

Figura 10-10
Riepilogo di categorizzazione per Anni all'attuale indirizzo

Intervallo	Punto finale		Numero di casi per livello di Precedenti insolvenze		
	Inferiore	Superiore	No	Sì	Totale
1	1[a]	7	1652	829	2481
2	7	14	1184	313	1497
3	14		908	114	1022
Totale			3744	1256	5000

Ciascun intervallo viene calcolato come Inferiore <= Anni di residenza all'indirizzo attuale < Superiore.

a. Illimitato

Il riepilogo per *Anni all'attuale indirizzo* mostra uno schema analogo. Come prevedibile in base alle statistiche dell'entropia di modello, le differenze tra intervalli in termini di soggetti inadempienti sono più nette per *Anni di permanenza nell'impiego attuale* che per *Anni all'attuale indirizzo*.

Intervallo	**Proporzione di inadempienti**
1	0.334
2	0.209
3	0.112

Figura 10-11
Riepilogo di categorizzazione per $$$Credit card debt in thousands

Bin	Punto finale		Numero di casi per livello di Previously defaulted		
	Inferiore	Superiore	No	Yes	Totale
1	[a]	,97	2169	466	2635
2	,97	1,91	848	307	1155
3	1,91	6,05	643	352	995
4	6,05	[a]	84	131	215
Totale			3744	1256	5000

Ciascun bin viene calcolato come Inferiore <= Credit card debt in thousands < Superiore.

a. Illimitato

Il riepilogo per *$$$Credit card debt in thousands* mostra lo schema inverso, con proporzioni di soggetti inadempienti crescenti man mano che aumentano i numeri di intervalli. *Anni di permanenza nell'impiego attuale* e *Anni all'attuale indirizzo* sembrano identificare meglio i soggetti con elevata probabilità di non inadempienza mentre *$$$Credit card debt in thousands* identifica meglio i soggetti con elevata probabilità di inadempienza.

Intervallo	**Proporzione di inadempienti**
1	0.177
2	0.266
3	0.354
4	0.609

Figura 10-12
Riepilogo di categorizzazione per Rapporto debito/reddito (x100)

	Punto finale		Numero di casi per livello di Previously defaulted		
Bin	Inferiore	Superiore	No	Yes	Totale
1	[a]	4,39	912	88	1000
2	4,39	12,09	2006	437	2443
3	12,09	18,71	625	386	1011
4	18,71	31,00	198	303	501
5	31,00	[a]	3	42	45
Totale			3744	1256	5000

Ciascun bin viene calcolato come Inferiore <= Debt to income ratio (x100) < Superiore.

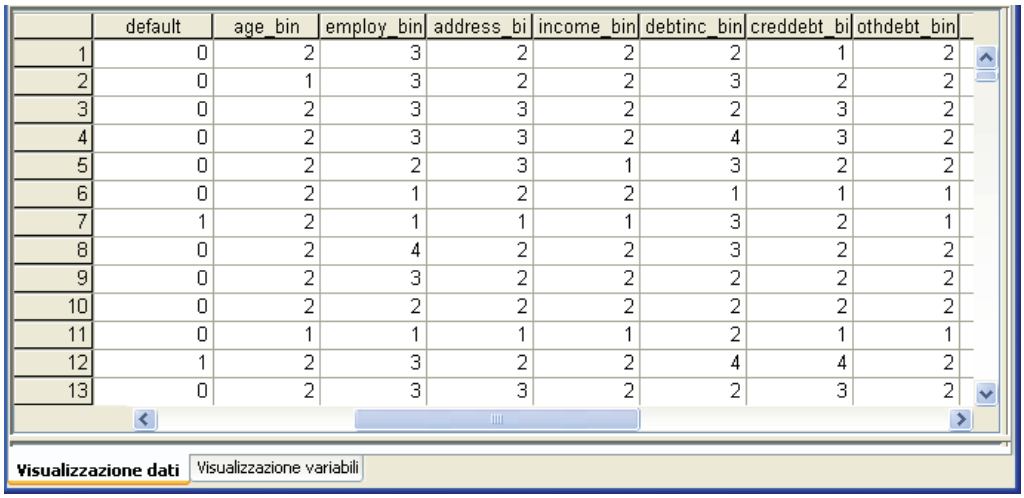
a. Illimitato

Il riepilogo per *Rapporto debito/reddito (x100)* mostra uno schema simile a*$$$Credit card debt in thousands*. Questa variabile presenta il valore di entropia di modello più basso ed è quindi il predittore migliore della probabilità di inadempienza. Classifica i probabili soggetti con elevata probabilità di inadempienza meglio di *Credit card debt in thousands* e classifica i soggetti con bassa probabilità di inadempienza quasi altrettanto bene di *Anni di permanenza nell'impiego attuale*.

Intervallo	**Proporzione di inadempienti**
1	0.088
2	0.179
3	0.382
4	0.605
5	0.933

Variabili categorizzate

Figura 10-13
Variabili categorizzate per bankloan_binning.sav nell'Editor dei dati



	default	age_bin	employ_bin	address_bi	income_bin	debtinc_bin	creddebt_bi	othdebt_bin
1	0	2	3	2	2	2	1	2
2	0	1	3	2	2	3	2	2
3	0	2	3	3	2	2	3	2
4	0	2	3	3	2	4	3	2
5	0	2	2	3	1	3	2	2
6	0	2	1	2	2	1	1	1
7	1	2	1	1	1	3	2	1
8	0	2	4	2	2	3	2	2
9	0	2	3	2	2	2	2	2
10	0	2	2	2	2	2	2	2
11	0	1	1	1	1	2	1	1
12	1	2	3	2	2	4	4	2
13	0	2	3	3	2	2	3	2

Visualizzazione dati | Visualizzazione variabili

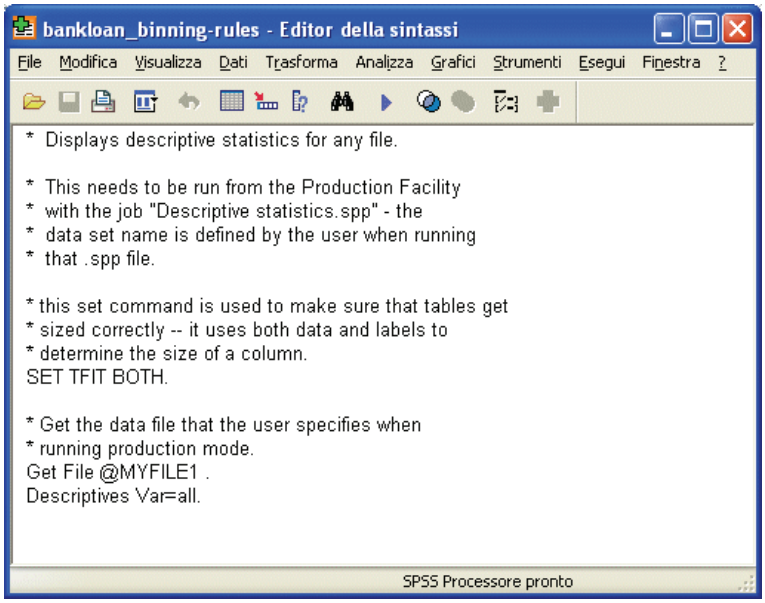
I risultati del processo di categorizzazione in questo insieme di dati sono evidenti nell'Editor dei dati. Queste variabili categorizzate risultano utili se si desidera produrre riepiloghi personalizzati dei risultati di categorizzazione mediante procedure descrittive o di reporting ma non è consigliabile utilizzare questo insieme di dati per creare un modello predittivo perché le regole di categorizzazione sono state generate utilizzando questi casi. Un piano migliore consiste nell'applicare le regole di categorizzazione a un altro insieme di dati contenente informazioni su altri clienti.

Applicazione delle regole di categorizzazione contenute nella sintassi

Durante l'esecuzione della procedura Categorizzazione ottimale è stato chiesto di salvare le regole di categorizzazione generate dalla procedura sotto forma di sintassi di comando.

- Aprire *bankloan_binning-rules.sps*.

Figura 10-14
File delle regole di sintassi



Per ogni variabile di input di categorizzazione esiste un blocco di sintassi di comandi che esegue la categorizzazione, imposta l'etichetta, il formato e il livello della variabile e imposta le etichette di valore per gli intervalli. Questi comandi possono essere applicati a un insieme di dati con le stesse variabili di *bankloan_binning.sav*.

- Aprire il file *bankloan.sav*. Per ulteriori informazioni, vedere l'argomento File di esempio in l'appendice A a pag. 137.
- Tornare alla vista dell'Editor della sintassi di *bankloan_binning-rules.sps*.

▶ Per applicare le regole di categorizzazione, dai menu dell'Editor della sintassi scegliere:
Esegui > Tutto...

Figura 10-15
Variabili categorizzate per bankloan.sav nell'Editor dei dati

	preddef3	age_bin	employ_bin	address_bin	income_bin	debtinc_bin	creddebt_bin	othdebt_bin
1	.21304	2	3	2	2	2	4	2
2	.43690	1	3	1	2	3	2	2
3	.14102	2	3	3	2	2	1	1
4	.10442	2	3	3	2	1	3	1
5	.43690	1	1	1	2	3	2	2
6	.23358	2	2	1	1	2	1	1
7	.81709	2	4	2	2	4	3	2
8	.11336	2	3	2	2	1	1	1
9	.66390	1	2	1	1	4	2	2
10	.51553	2	1	2	1	4	3	1
11	.09055	1	1	1	1	1	1	1
12	.13631	1	2	1	1	2	1	1
13	.22890	2	4	3	2	2	3	2
14	.40484	2	2	2	2	3	2	2
15	.20866	2	4	3	2	2	3	2

Visualizzazione dati | Visualizzazione variabili

Le variabili contenute in *bankloan.sav* sono state categorizzate in base alle regole generate dall'esecuzione della procedura Categorizzazione ottimale in *bankloan_binning.sav*. Questo insieme di dati è ora pronto per essere utilizzato nella creazione di modelli predittivi che preferiscono o richiedono variabili categoriali.

Riepilogo

La procedura Categorizzazione ottimale ha consentito di generare regole di categorizzazione per variabili di scala che si configurano come potenziali predittori della probabilità di inadempienza e le regole così generate sono state applicate a un insieme di dati distinto.

Durante il processo di categorizzazione si è notato che le variabili categorizzate *Anni di permanenza nell'impiego attuale* e *Anni all'attuale indirizzo* identificano meglio i soggetti con un'elevata probabilità di non inadempienza mentre *Credit card debt in thousands* identifica meglio i soggetti con un'elevata probabilità di inadempienza. Questa interessante osservazione fornisce ulteriori indicazioni, utili per la creazione di modelli predittivi della probabilità di inadempienza. Se l'interesse principale consiste nell'evitare l'inadempienza, *$$$Credit card debt in thousands* avrà maggiore importanza di *Anni di permanenza nell'impiego attuale* e *Anni all'attuale indirizzo*. Se la priorità è ampliare la base clienti, *Anni di permanenza nell'impiego attuale* e *Anni all'attuale indirizzo* saranno più importanti.

Appendice

A

File di esempio

Il file di esempio installato con il prodotto si trova nella sottodirectory *Samples* della directory di installazione. La sottodirectory Samples contiene cartelle separate per ciascuna delle seguenti lingue: Inglese, Francese, Tedesco, Italiano, Giapponese, Coreano, Polacco, Russo, Cinese semplificato, Spagnolo e Cinese tradizionale.

Non tutti i file di esempio sono disponibili in tutte le lingue. Se un file di esempio non è disponibile in una lingua, la cartella di tale lingua contiene una versione inglese del file.

Descrizioni

Questa sezione contiene brevi descrizioni dei file di esempio utilizzati negli esempi riportati in tutta la documentazione.

- **accidents.sav.** File di dati ipotetici che prende in esame una compagnia di assicurazioni impegnata nello studio dei fattori di rischio correlati all'età e al sesso per gli incidenti automobilistici che si verificano in una determinata regione. Ciascun caso corrisponde a una classificazione incrociata della categoria relativa età e del sesso.
- **adl.sav.** File di dati ipotetici che prende in esame l'impegno richiesto per determinare i vantaggi di un tipo di terapia proposto per i pazienti con problemi di cuore. I medici hanno assegnato in modo casuale i pazienti con problemi di cuore di sesso femminile a uno di due gruppi. Al primo gruppo è stata assegnata la terapia fisica standard; al secondo gruppo, un'ulteriore terapia di supporto psicologico. Dopo tre mesi di trattamenti, a ciascuna capacità dei pazienti che consente di riprendere le normali attività giornaliere è stato assegnato un punteggio come variabile ordinale.
- **advert.sav.** File di dati ipotetici che prende in esame l'impegno di un rivenditore al dettaglio che desidera esaminare la relazione tra il denaro speso per la pubblicità e le vendite risultanti. Finora sono stati raccolti i dati delle vendite precedenti e i relativi costi pubblicitari.
- **aflatoxin.sav.** File di dati ipotetici che prende in esame il test di raccolti di mais con presenza di Aflatossina, un veleno la cui concentrazione varia notevolmente nei raccolti. Una macchina per la lavorazione dei cereali ha ricevuto 16 campioni da ciascuno degli otto raccolti di mais e ha misurato i livelli di Aflatossina in parti per miliardo (PPB).
- **aflatoxin20.sav.** Questo file di dati contiene le misurazioni di Aflatossina di ciascuno dei 16 campioni di quattro raccolti e otto campioni dal file di dati *aflatoxin.sav*.
- **anorectic.sav.** Per trovare una sintomatologia standardizzata del comportamento anoressico/bulimico, i ricercatori (Van der Ham, Meulman, Van Strien, e Van Engeland, 1997) hanno condotto uno studio basato su 55 adolescenti affetti da disordini alimentari conosciuti. Ogni paziente è stato visitato quattro volte in quattro anni, per un totale di 220 visite. Durante ogni visita, ai pazienti sono stati assegnati punteggi per ciascuno dei 16 sintomi. I punteggi

relativi ai sintomi sono assenti per il paziente 71 alla visita 2, il paziente 76 alla visita 2 e il paziente 47 alla visita 3, con 217 osservazioni valide.

- **autoaccidents.sav.** File di dati ipotetici che prende in esame l'impegno di un analista che opera nel campo delle assicurazioni per creare un modello del numero di incidenti automobilistici per conducente. Il modello prende in esame anche l'età e il sesso del conducente. Ciascun caso rappresenta un diverso conducente e riporta il sesso e l'età (in anni) del conducente e il numero di incidenti automobilistici negli ultimi cinque anni.
- **band.sav.** Questo file di dati ipotetici contiene le cifre sulle vendite settimanali di CD conseguite da un gruppo musicale. Il file include anche i dati di tre possibili variabili predittore.
- **bankloan.sav.** File di dati ipotetici che prende in esame l'impegno di una banca nel tentativo di ridurre il tasso di inadempienza nel rimborso di un prestito. Il file contiene informazioni finanziarie e demografiche su 850 vecchi e potenziali clienti. I primi 700 casi riguardano i clienti a cui sono stati concessi dei prestiti precedentemente. Gli ultimi 150 casi riguardano i potenziali clienti che la banca deve classificare come rischi di credito positivi o negativi.
- **bankloan_binning.sav.** File di dati ipotetici che contiene informazioni finanziarie e demografiche su 5000 vecchi clienti.
- **behavior.sav.** In un classico esempio (Prezzo e Bouffard, 1974), è stato chiesto a 52 studenti di classificare una combinazione di 15 situazioni e 15 comportamenti utilizzando una scala da 0="molto appropriato" a 9="molto inadeguato". I valori medi riferiti ai partecipanti sono stati considerati dissimilarità.
- **behavior_ini.sav.** Questo file di dati contiene la configurazione iniziale di una soluzione a due dimensioni per *behavior.sav*.
- **brakes.sav.** File di dati ipotetici che prende in esame il controllo di qualità di un'industria che produce freni a disco per automobili con elevate prestazioni. Il file di dati contiene le misurazioni del diametro di 16 dischi da ciascuna delle otto macchine di produzione. L'obiettivo finale è ottenere un diametro dei dischi pari a 322 millimetri.
- **breakfast.sav.** In uno studio classico (Green e Rao, 1972), è stato chiesto a 21 studenti MBA della Wharton School e ai loro consorti di classificare 15 cibi da colazione in ordine di preferenza, dove il valore 1 corrispondeva all'alimento preferito in assoluto e il valore 15 a quello meno preferito. Le loro preferenze sono state registrate per sei diversi scenari, che comprendevano tutti gli scenari compresi tra "Preferenza generale" e "Solo snack con bibita".
- **breakfast-overall.sav.** Questo file contiene le preferenze degli alimenti della colazione solo per il primo scenario, "Preferenza generale".
- **broadband_1.sav.** File di dati ipotetici che contiene il numero di sottoscrittori, per area, di un provider di servizi a banda larga nazionale. Il file di dati contiene il numero dei sottoscrittori mensili di 85 aree in un periodo di quattro anni.
- **broadband_2.sav.** Questo file è identico al file *broadband_1.sav*, ma contiene i dati per ulteriori tre mesi.
- **car_insurance_claims.sav.** Un insieme di dati presentato e analizzato altrove (McCullagh e Nelder, 1989) riguarda le richieste di risarcimento auto. La quantità media di richieste di risarcimento può essere adattata come avente una distribuzione gamma, utilizzando una funzione di collegamento inverso per correlare la media della variabile dipendente a una

combinazione lineare di età del contraente della polizza e tipo e anni del veicolo. Il numero delle richieste di risarcimento specificato può essere utilizzato come peso scalato.

- **car_sales.sav.** Questo file di dati ipotetici contiene le stime sulle vendite, i prezzi di listino e le specifiche fisiche di numerose marche e modelli di veicoli. I prezzi di listino e le specifiche fisiche sono state ottenute dal sito *edmunds.com* e dai siti dei produttori.
- **car_sales_uprepared.sav.** Questa è una versione modificata di *car_sales.sav* che non comprende versioni trasformate dei campi.
- **carpet.sav.** Come esempio tipico (Green e Wind, 1973), un'azienda interessata alla commercializzazione di un nuovo battitappeto desidera esaminare l'influenza di cinque fattori sulle preferenze del consumatore, ovvero design della confezione, marca, prezzo, la presenza di un *marchio di qualità* e una garanzia "Soddisfatti o rimborsati". Esistono tre livelli di fattore per il design della confezione, che differiscono per la posizione della spazzola dell'applicatore; tre marchi (*K2R*, *Glory* e *Bissell*); tre livelli di prezzo e due livelli (no o sì) per ciascuno degli ultimi due fattori. Dieci consumatori sono classificati in 22 profili definiti da questi fattori. La variabile *Preferenza* include il rango delle classificazioni medie per ogni profilo. Classificazioni basse corrispondono a una preferenza elevata. La variabile riflette una misura globale della preferenza per ogni profilo.
- **carpet_prefs.sav.** Questo file di dati si basa sullo stesso esempio del file *carpet.sav*, ma contiene le classificazioni effettive raccolte da ciascuno dei 10 clienti. Ai clienti è stato chiesto di classificare 22 profili di prodotti in ordine di preferenza. Le variabili da *PREF1* a *PREF22* contengono gli ID dei profili associati, come definito nel file *carpet_plan.sav*.
- **catalog.sav.** File di dati ipotetico che contiene le cifre sulle vendite mensili di tre prodotti venduti da una società di vendita per corrispondenza. Il file include anche i dati di cinque possibili variabili predittore.
- **catalog_seasfac.sav.** Questo file di dati è uguale al file *catalog.sav* con l'eccezione che contiene un insieme di fattori stagionali calcolati dalla procedura Decomposizionale stagionale insieme a variabili di dati.
- **cellular.sav.** File di dati ipotetici che prende in esame l'impegno di un'azienda di telefonia cellulare nel tentativo di ridurre il churn, ovvero l'abbandono dei clienti. Agli account vengono applicati i punteggi relativi alla propensione al churn, con valori compresi tra 0 e 100. Gli account con punteggio pari a 50 o superiore è probabile che stiano cercando nuovi provider.
- **ceramics.sav.** File di dati ipotetici che prende in esame l'impegno di un produttore che desidera stabilire se una nuova lega premium ha una maggiore resistenza al calore rispetto alla lega standard. Ciascun caso rappresenta il test separato di una delle leghe. È indicata la temperatura massima alla quale può essere sottoposto il cuscinetto.
- **cereal.sav.** File di dati ipotetici che prende in esame le preferenze relative agli alimenti della colazione di un campione di 880 persone. Il file riporta anche l'età, il sesso e lo stato civile del campione e se le persone conducono uno stile di vita attivo (in base a un'attività sportiva con frequenza di due volte alla settimana). Ogni caso rappresenta un rispondente separato.
- **clothing_defects.sav.** File di dati ipotetici che prende in esame il processo di controllo di qualità di un'industria di abbigliamento. Per ciascun lotto prodotto nella fabbrica, gli ispettori prelevano un campione di abiti per contare il numero dei capi che non sono accettabili per la vendita.

- **coffee.sav.** Questo file di dati contiene informazioni sulle immagini percepite di sei marche di caffè freddo (Kennedy, Riquier, e Sharp, 1996). Per ciascuno dei 23 attributi dell'immagine del caffè freddo, sono state selezionate tutte le marche descritte da tale attributo. Le sei marche sono indicate dalle sigle AA, BB, CC, DD, EE e FF per tutelare la confidenzialità dei dati.
- **contacts.sav.** File di dati ipotetici che prende in esame l'elenco dei contatti di un gruppo di rappresentanti di vendita di computer aziendali. Ciascun contatto è classificato in base al reparto della società in cui lavora e dalle relative categorie aziendali. Il file riporta anche l'importo dell'ultima vendita effettuata, il tempo trascorso dall'ultima vendita e le dimensioni della società del contatto.
- **creditpromo.sav.** File di dati ipotetici che prende in esame l'impegno di un grande magazzino nel tentativo di valutare l'efficacia di una recente promozione con carta di credito. A tale scopo, sono stati selezionati 500 titolari di carta in modo casuale. Alla metà di questi è stato inviato un annuncio promozionale che comunica la riduzione del tasso d'interesse nel caso di acquisti effettuati entro i tre mesi successivi. All'altra metà è stato inviato un annuncio stagionale standard.
- **customer_dbase.sav.** File di dati ipotetico che prende in esame l'impegno di una società nel tentativo di utilizzare le informazioni contenute nel proprio database dei dati per creare offerte speciali per i clienti che più probabilmente risponderanno all'offerta. È stato selezionato in modo casuale un sottoinsieme della base dei clienti a cui è stata inviata l'offerta speciale e sono state registrate le risposte ricevute.
- **customer_information.sav.** File di dati ipotetici contenente le informazioni postali del cliente, ad esempio il nome e l'indirizzo.
- **customer_subset.sav.** Un sottoinsieme di 80 casi da *customer_dbase.sav*.
- **customers_model.sav.** File di dati ipotetici che contiene il nominativo delle persone a cui è stata inviata una campagna di marketing. I dati includono informazioni demografiche, un riepilogo della cronologia degli acquisti e se ciascuna persona ha risposto alla campagna. Ogni caso rappresenta una persona separata.
- **customers_new.sav.** File di dati ipotetici che contiene i nominativi delle persone che sono state evidenziate come potenziali candidati per una campagna di marketing. I dati includono informazioni demografiche e un riepilogo sulla cronologia degli acquisti di ciascuna persona. Ogni caso rappresenta una persona separata.
- **debate.sav.** File di dati ipotetici che prende in esame le risposte appaiate a un'indagine da parte dei partecipanti a un dibattito politico prima e dopo il dibattito. Ogni caso rappresenta un rispondente separato.
- **debate_aggregate.sav.** File di dati ipotetici che aggrega le risposte contenute nel file *debate.sav*. Ciascun caso corrisponde a una classificazione incrociata della preferenza prima e dopo il dibattito.
- **demo.sav.** File di dati ipotetici che prende in esame un database di clienti che hanno fatto acquisti al fine di inviare offerte mensili tramite il metodo del direct mailing. Viene registrata la risposta dei clienti, sia che abbiano aderito all'offerta o meno, insieme a diverse informazioni demografiche.
- **demo_cs_1.sav.** File di dati ipotetici che prende in esame il primo passo che una società intraprende per compilare un database con informazioni ricavate dai sondaggi. Ogni caso rappresenta una diversa città. Sono registrate anche le informazioni sulla regione, provincia, distretto e città.

- **demo_cs_2.sav.** File di dati ipotetici che prende in esame il secondo passo che una società intraprende per compilare un database con informazioni ricavate dai sondaggi. Ogni caso rappresenta una diversa unità di abitazione, ricavata dalle città selezionate nel primo passo. Sono registrate anche le informazioni sulla regione, provincia, distretto, città, suddivisione e unità. Il file include inoltre informazioni sul campionamento ottenute dai primi due stadi del disegno.
- **demo_cs.sav.** File di dati ipotetici che contiene informazioni sulle indagini raccolte utilizzando un disegno di campionamento complesso. Ogni caso rappresenta una diversa unità di abitazione. Sono registrate diverse informazioni demografiche e sul campionamento.
- **dmdata.sav.** File di dati ipotetici che contiene informazioni demografiche e di acquisto di una società di direct marketing. *dmdata2.sav* contiene informazioni su un sottoinsieme di contatti che hanno ricevuto un mailing di prova e *dmdata3.sav* contiene informazioni sui contatti rimanenti che non hanno ricevuto il mailing di prova.
- **dietstudy.sav.** File di dati ipotetici che contiene il risultato di uno studio ipotetico sulla dieta chiamato "Stillman diet" (Rickman, Mitchell, Dingman, e Dalen, 1974). Ogni caso rappresenta un diverso soggetto e ne riporta il peso prima e dopo la dieta in libbre e i livelli dei trigliceridi in mg/100 ml.
- **dvdplayer.sav.** File di dati ipotetici che prende in esame lo sviluppo di un nuovo lettore DVD. Utilizzando un prototipo, il personale addetto al marketing ha raccolto dati sui gruppi di interesse. Ogni caso rappresenta un diverso utente che è stato sottoposto all'indagine e include informazioni demografiche personali dell'utente e sulle risposte che ha fornito riguardo al prototipo.
- **german_credit.sav.** Questo file di dati contiene informazioni ricavate dall'insieme di dati "German Credit" del Repository of Machine Learning Databases (Blake e Merz, 1998) presso la University of California, Irvine.
- **grocery_1month.sav.** Questo file di dati ipotetici corrisponde al file di dati *grocery_coupons.sav* con gli acquisti settimanali organizzati in modo che ogni caso corrisponda a un cliente separato. Alcune delle variabili che cambiano settimanalmente non vengono riportate nei risultati; l'importo speso registrato corrisponde ora alla somma degli importi spesi durante le quattro settimane dello studio.
- **grocery_coupons.sav.** File di dati ipotetici che contiene i dati sui sondaggi raccolti da una catena di drogherie interessata alle abitudini di acquisto dei suoi clienti. Ciascun cliente viene seguito per quattro settimane e ciascun caso corrisponde a una settimana per cliente con informazioni sul luogo degli acquisti e i tipi di acquisti, incluso l'importo speso nelle drogherie durante la settimana.
- **guttman.sav.** Bell (Bell, 1961) ha presentato una tabella per illustrare i possibili gruppi sociali. Guttman (Guttman, 1968) ha utilizzato una parte di tale tabella, in cui cinque variabili che descrivono elementi come l'interazione sociale, i sentimenti di appartenenza a un gruppo, la vicinanza fisica dei membri e il grado di formalità della relazione, sono state incrociate con cinque gruppi sociali teorici, compresi folla (ad esempio, le persone presenti a una partita di calcio), uditorio (ad esempio, di uno spettacolo teatrale o di una lezione universitaria), pubblico (ad esempio televisivo), calca (come una folla, ma con un'interazione molto maggiore), gruppi primari (intimi), gruppi secondari (volontari) e la comunità moderna (unione non stretta derivante da una vicinanza fisica elevata e dall'esigenza di servizi specializzati).

- **health_funding.sav.** File di dati ipotetici che contiene i dati sui fondi di assistenza sanitaria (importo per 100 persone), sui tassi di malattie (tasso per 10.000 persone) e sulle visite ai fornitori di assistenza sanitaria (tasso per 10.000 persone). Ogni caso rappresenta una diversa città.
- **hivassay.sav.** File di dati ipotetici che prende in esame l'impegno di un'industria farmaceutica nel tentativo di sviluppare un'analisi che riesca a rilevare in tempi brevi l'infezione da virus HIV. I risultati dell'analisi sono otto sfumature di colore rosso sempre più intenso; le sfumature più intense indicano la maggiore probabilità di infezione. Un esperimento di laboratorio è stato condotto su 2000 campioni di sangue. La metà di questi è risultata infetta al virus HIV, l'altra metà non è risultata infetta.
- **hourlywagedata.sav.** File di dati ipotetici che prende in esame la paga oraria degli infermieri occupati presso uffici e ospedali e in base ai diversi livelli di esperienza.
- **insurance_claims.sav.** File di dati ipotetici che prende in esame una compagnia di assicurazioni impegnata nella creazione di un modello per contrassegnare le richieste di risarcimento sospette e potenzialmente fraudolente. Ogni caso rappresenta una richiesta di risarcimento separata.
- **insure.sav.** File di dati ipotetici che prende in esame una compagnia di assicurazioni impegnata nello studio dei fattori di rischio, che indicano l'eventualità che un cliente presenti una domanda di indennizzo in un contratto assicurativo sulla vita della durata di dieci anni. Ogni caso nel file di dati rappresenta una coppia di contratti. In un contratto sono contenute informazioni su una richiesta di risarcimento, l'altro sull'età e sul sesso.
- **judges.sav.** File di dati ipotetici che prende in esame il punteggio assegnato, da giurie qualificate (più un appassionato) a 300 prestazioni sportive. Ciascuna riga rappresenta una diversa prestazione; i giudici hanno esaminato le stesse prestazioni.
- **kinship_dat.sav.** Rosenberg e Kim (Rosenberg e Kim, 1975) si prefiggono di analizzare 15 termini indicanti parentela (zia, fratello, cugino, padre, nipote femmina, di nonni, nonno, nonna, nipote maschio di nonni, madre, nipote maschio di zii), nipote femmina di zii, sorella, figlio, zio). Hanno richiesto a quattro gruppi di studenti universitari (due composti da femmine e due da maschi) di ordinare questi termini in base alla similiarità. A due gruppi (uno femminile e uno maschile) è stato richiesto di effettuare l'ordinamento due volte, con il secondo ordinamento basato su un criterio diverso rispetto al primo. Di conseguenza, sono state ottenute sei "sorgenti" in totale. Ogni sorgente corrisponde a una matrice di prossimità 15×15, le cui celle sono uguali al numero delle persone in una sorgente meno il numero di volte in cui gli oggetti sono stati ripartiti insieme nella sorgente.
- **kinship_ini.sav.** Questo file di dati contiene la configurazione iniziale di una soluzione a tre dimensioni per *kinship_dat.sav*.
- **kinship_var.sav.** Questo file di dati contiene variabili indipendenti relative a *sesso*, *generazione* e *grado* di separazione che possono essere utilizzate per interpretare le dimensioni di una soluzione per *kinship_dat.sav*. In modo specifico, tali variabili possono essere utilizzate per limitare lo spazio della soluzione a una combinazione lineare di tali variabili.
- **marketvalues.sav.** File di dati che prende in esame le vendite di abitazioni in un nuovo centro abitato in Algonquin, Ill., durate gli anni 1999–2000. Tali vendite sono una questione di dominio pubblico.

- **nhis2000_subset.sav.** Il National Health Interview Survey (NHIS) è un sondaggio di grandi dimensioni condotto sulla popolazione civile americana. Le interviste vengono realizzate di persona e si basano su un campione rappresentativo di famiglie a livello nazionale. Per ogni membro di una famiglia vengono raccolte osservazioni e informazioni di carattere demografico relative allo stato di salute. Questo file di dati contiene un sottoinsieme delle informazioni ottenute dall'indagine del 2000. National Center for Health Statistics. National Health Interview Survey, 2000. File di dati e documentazione di dominio pubblico. *ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHIS/2000/*. Accesso 2003.
- **ozone.sav** I dati includono 330 osservazioni basate su sei variabili meteorologiche per quantificare la concentrazione dell'ozono dalle variabili rimanenti. I precedenti ricercatori, (Breiman e Friedman, 1985) e (Hastie e Tibshirani, 1990), hanno rilevato non linearità tra queste variabili, che impediscono un approccio di regressione standard.
- **pain_medication.sav.** File di dati ipotetici che contiene i risultati di un test clinico per stabilire la cura antinfiammatoria per il trattamento del dolore generato dall'artrite cronica. Di particolare interesse, il test ha evidenziato il tempo che impiega il farmaco ad avere effetto e il confronto con altri farmaci esistenti.
- **patient_los.sav.** File di dati ipotetici che contiene informazioni sul trattamento dei pazienti ricoverati per sospetto di infarto del miocardio. Ogni caso corrisponde a un diverso paziente e contiene diverse variabili correlate alla degenza nell'ospedale.
- **patlos_sample.sav.** File di dati ipotetici che contiene informazioni sul trattamento di un campione di pazienti curato con trombolitici durante la degenza per infarto del miocardio. Ogni caso corrisponde a un diverso paziente e contiene diverse variabili correlate alla degenza nell'ospedale.
- **polishing.sav.** File di dati "Nambeware Polishing Times" di Data and Story Library. Prende in esame l'impegno di un'industria di stoviglie in metallo (Nambe Mills, Santa Fe, N. M.) nel tentativo di pianificare il proprio piano di produzione. Ogni caso rappresenta un diverso articolo nella linea dei prodotti. Per ciascun articolo sono indicati il diametro, il tempo di lucidatura, il prezzo e il tipo di prodotto.
- **poll_cs.sav.** File di dati ipotetici che prende in esame i sondaggi per stabilire il livello di sostegno pubblico nei confronti di un disegno di legge prima che diventi una legge vera e propria. I casi corrispondono ai votanti registrati. Ciascun caso riporta informazioni sulla contea, sul comune e sul quartiere in cui vive il votante.
- **poll_cs_sample.sav.** File di dati ipotetici che contiene un campione dei votanti elencati nel file *poll_cs.sav*. Il campione è stato selezionato in base al disegno specificato nel file di piano *poll.csplan* e questo file di dati contiene le probabilità di inclusione e i pesi del campione. Tuttavia, notare che poiché fa uso del metodo PPS (probability-proportional-to-size, probabilità proporzionale alla dimensione), esiste anche un file contenente le probabilità di selezione congiunte (*poll_jointprob.sav*). Le ulteriori variabili corrispondenti ai dati demografici dei votanti e alla loro opinione sul disegno di legge, sono state raccolte e aggiunte al file di dati dopo aver acquisito il campione.
- **property_assess.sav.** File di dati ipotetici che prende in esame l'impegno di un perito di una contea nel tentativo di mantenere gli accertamenti sui valori delle proprietà aggiornati in base alle risorse limitate. I casi rappresentano le proprietà vendute nella contea nello scorso anno. Ogni caso nel file di dati contiene informazioni sul comune in cui si trova la proprietà, il perito che per ultimo ha visitato la proprietà, il tempo trascorso dall'accertamento, la valutazione fatta in tale momento e il valore di vendita della proprietà.

- **property_assess_cs.sav.** File di dati ipotetici che prende in esame l'impegno di un perito di uno stato nel tentativo di mantenere aggiornati gli accertamenti sui valori delle proprietà in base alle risorse limitate. I casi corrispondono alle proprietà nello stato. Ogni caso nel file di dati include informazioni sulla contea, il comune e il quartiere in cui risiede la proprietà, la data dell'ultimo accertamento e la valutazione fatta in tale data.
- **property_assess_cs_sample.sav.** File di dati ipotetici che contiene un campione delle proprietà elencate nel file *property_assess_cs.sav*. Il campione è stato selezionato in base al disegno specificato nel file di piano *property_assess.csplan* e questo file di dati contiene le probabilità di inclusione e i pesi del campione. L'ulteriore variabile *Valore corrente* è stata raccolta e aggiunta al file di dati dopo aver acquisito il campione.
- **recidivism.sav.** File di dati ipotetici che prende in esame l'impegno delle Forze dell'Ordine nel tentativo di valutare il tasso di recidività nella propria area di giurisdizione. Ogni caso corrisponde a un precedente trasgressore e include le informazioni demografiche, alcuni dettagli sul primo crimine, il tempo trascorso fino al secondo arresto e se tale arresto è avvenuto entro due anni dal primo.
- **recidivism_cs_sample.sav.** File di dati ipotetici che prende in esame l'impegno delle Forze dell'Ordine nel tentativo di valutare il tasso di recidività nella propria area di giurisdizione. Ogni caso corrisponde a un trasgressore precedente, rilasciato dopo il primo arresto durante il mese di giugno del 2003 e registra le relative informazioni demografiche, alcuni dettagli sul primo crimine commesso e i dati del secondo arresto, se si è verificato prima della fine di giugno del 2006. I trasgressori sono stati selezionati dai dipartimenti sottoposti a campione in base al piano di campionamento specificato nel file *recidivism_cs.csplan*. Poiché viene utilizzato un metodo PPS (Probability-Proportional-to-Size, probabilità proporzionale alla dimensione), esiste anche un file contenente le probabilità di selezione congiunte (*recidivism_cs_jointprob.sav*).
- **rfm_transactions.sav.** File di dati ipotetici contenente i dati delle transazioni di acquisto, inclusa la data di acquisto, gli articoli acquistati e il valore monetario di ciascuna transazione.
- **salesperformance.sav.** File di dati ipotetici che prende in esame la valutazione di due nuovi corsi di formazione alle vendite. Sessanta dipendenti, divisi in tre gruppi, ricevono tutti la formazione standard. In più, al gruppo 2 viene assegnato un corso di formazione tecnica e al gruppo 3 un'esercitazione pratica. Alla fine del corso di formazione, ciascun dipendente viene sottoposto a un esame e il punteggio conseguito viene registrato. Ciascun caso nel file di dati rappresenta un diverso partecipante. Il file di dati include il gruppo a cui è assegnato il partecipante e il punteggio conseguito all'esame finale.
- **satisf.sav.** File di dati ipotetico che prende in esame un'indagine sulla soddisfazione dei clienti condotta da una società di vendita al dettaglio presso 4 negozi. Sono stati intervistati 582 clienti e ciascun caso rappresenta le risposte ottenute da un singolo cliente.
- **screws.sav.** Questo file di dati contiene informazioni sulle caratteristiche di viti, bulloni, dadi e puntine (Hartigan, 1975).
- **shampoo_ph.sav.** File di dati ipotetici che prende in esame il processo di controllo di qualità di un'industria di prodotti per capelli. A intervalli di tempo regolari, vengono misurati sei diversi lotti prodotti e ne viene registrato il relativo pH. I valori accettati sono compresi tra 4,5 e 5,5.
- **ships.sav.** Ad esempio, un insieme di dati presentato e analizzato altrove (McCullagh et al., 1989) riguarda i danni subiti dalle navi da carico a causa delle onde. I conteggi degli incidenti possono essere presentati con un tasso di Poisson in base al tipo di nave, al periodo di

costruzione e al periodo di servizio. I mesi di servizio aggregati di ciascuna cella della tabella generata dalla classificazione incrociata dei fattori fornisce i valori di esposizione al rischio.

- **site.sav.** File di dati ipotetici che prende in esame l'impegno di una società nella scelta di nuovi siti in cui espandere la propria presenza. La società ha incaricato due consulenti separati che, oltre a valutare i siti e presentare un report completo, devono classificarli come potenzialmente "molto adatti", "adatti" o "poco adatti".
- **smokers.sav.** Questo file di dati è un estratto del 1998 National Household Survey of Drug Abuse e rappresenta un campione probabile di famiglie americane. (*http://dx.doi.org/10.3886/ICPSR02934*) Il primo passo nell'analisi di questo file di dati consiste quindi nel pesare i dati per rispecchiare le tendenze della popolazione.
- **stroke_clean.sav.** File di dati ipotetici che riporta lo stato di un database medico dopo averne eseguito la pulizia utilizzando le procedure del modulo Data Preparation.
- **stroke_invalid.sav.** File di dati ipotetici che riporta lo stato iniziale di un database medico e contiene numerosi errori di immissione dati.
- **stroke_survival.** Questo file di dati ipotetici riguarda i tempi di sopravvivenza per i pazienti che, dopo avere completato un programma riabilitativo in seguito a un ictus postischemico, affrontano alcune sfide. Dopo l'attacco, viene annotata l'occorrenza dell'infarto miocardiaco, dell'ictus ischemico o emorragico e viene registrata l'ora dell'evento. Questo campione viene troncato a sinistra perché include solo i pazienti che sono sopravvissuti fino alla fine del programma riabilitativo post-ictus.
- **stroke_valid.sav.** File di dati ipotetici che riporta lo stato di un database medico dopo il controllo dei valori eseguito con la procedura Convalida i dati. Il database contiene comunque casi potenzialmente anomali.
- **survey_sample.sav.** File di dati che contiene i dati dell'indagine, compresi i dati demografici e varie misure dell'atteggiamento. Si basa su un sottoinsieme di variabili tratte dal 1998 NORC General Social Survey, benché i valori di alcuni dati siano stati modificati e siano state aggiunte variabili fittizie a scopo dimostrativo.
- **telco.sav.** File di dati ipotetici che prende in esame l'impegno di un'azienda di telecomunicazioni nel tentativo di ridurre il churn, ovvero l'abbandono dei propri clienti. Ciascun caso rappresenta un cliente separato e riporta diverse informazioni demografiche e sull'uso del servizio.
- **telco_extra.sav.** Questo file di dati è simile al file *telco.sav*, ma le variabili "tenure" e spesa del cliente trasformata tramite logaritmo sono state sostituite dalle variabili di spesa del cliente trasformata tramite logaritmo standardizzate.
- **telco_missing.sav.** Questo file di dati è un sottoinsieme del file di dati *telco.sav*, ma alcuni dei valori di dati demografici sono stati sostituiti con valori mancanti.
- **testmarket.sav.** File di dati ipotetici che prende in esame i piani di una catena di fast food per aggiungere un nuovo prodotto al proprio menu. Sono previste tre campagne promozionali del nuovo prodotto. Il prodotto viene introdotto in diversi mercati selezionati in modo casuale. Per ogni sede viene utilizzata una promozione differente registrando le vendite settimanali della nuova voce per le prime quattro settimane. Ogni caso rappresenta un luogo e una settimana diversi.

- **testmarket_1month.sav.** Questo file di dati ipotetici corrisponde al file *testmarket.sav* con le vendite settimanali organizzate in modo che ogni caso corrisponda a un luogo separato. Alcune delle variabili che cambiano settimanalmente non vengono riportate nei risultati; le vendite registrate corrispondono ora alla somma delle vendite conseguite durante le quattro settimane dello studio.
- **tree_car.sav.** File di dati ipotetici che contiene dati demografici e sul prezzo di acquisto dei veicoli.
- **tree_credit.sav.** File di dati ipotetici che contiene dati demografici e sulla cronologia dei mutui di una banca.
- **tree_missing_data.sav.** File di dati ipotetici che contiene dati demografici e sulla cronologia dei mutui di una banca con un numero elevato di valori mancanti.
- **tree_score_car.sav.** File di dati ipotetici che contiene dati demografici e sul prezzo di acquisto dei veicoli.
- **tree_textdata.sav.** File di dati semplice con due variabili destinato principalmente per mostrare lo stato predefinito delle variabili prima dell'assegnazione dei livelli di misurazione e delle etichette dei valori.
- **tv-survey.sav.** File di dati ipotetici che prende in esame un sondaggio condotto da una emittente televisiva che deve stabilire se estendere la durata di un programma di successo. A un campione di 906 intervistati è stato chiesto se preferisce guardare il programma con diverse condizioni. Ciascuna riga rappresenta un diverso intervistato e ciascuna colonna una diversa condizione.
- **ulcer_recurrence.sav.** Questo file contiene informazioni parziali su uno studio svolto per mettere a confronto l'efficacia di due terapie preventive per la recidiva delle ulcere. Fornisce un ottimo esempio di dati acquisiti a intervalli ed è stato presentato e analizzato in altri luoghi (Collett, 2003).
- **ulcer_recurrence_recoded.sav.** In questo file sono contenute le informazioni del file *ulcer_recurrence.sav* riorganizzate per consentire di presentare la probabilità degli eventi per ciascun intervallo dello studio, anziché solo alla fine. È stato presentato e analizzato in altri luoghi (Collett et al., 2003).
- **verd1985.sav.** Questo file di dati prende in esame un'indagine (Verdegaal, 1985). Sono state registrate le risposte di quindici soggetti a otto variabili. Le variabili di interesse sono suddivise in tre insiemi. L'insieme 1 include *età* e *statociv*, l'insieme 2 include *andom* e *giornale* e l'insieme 3 include *musica* e *vicinato*. *Andom* viene scalata come nominale multipla ed *età* come ordinale; tutte le altre variabili vengono scalate come nominali singole.
- **virus.sav.** File di dati ipotetici che prende in esame l'impegno di un ISP (Internet Service Provider) nel tentativo di determinare gli effetti che un virus può generare nelle sue reti. Si è tenuta traccia della percentuale (approssimativa) di traffico e-mail infettato da virus sulla rete in un lasso di tempo, dal momento dell'individuazione fino alla soppressione della minaccia.
- **wheeze_steubenville.sav.** Questo file è un sottoinsieme di uno studio longitudinale degli effetti che l'inquinamento provoca sulla salute dei bambini (Ware, Dockery, Spiro III, Speizer, e Ferris Jr., 1984). I dati contengono misure binarie ripetute del livello di asma dei bambini

della città di Steubenville, Ohio, di 7, 8, 9 e 10 anni. I dati indicano anche se la mamma dei bambini era fumatrice durante il primo anno dello studio.

- **workprog.sav.** File di dati ipotetici che prende in esame un programma di lavoro governativo il cui obiettivo è fornire attività più adatte alle persone diversamente abili. È stato seguito un campione di potenziali partecipanti al programma, alcuni dei quali sono stati selezionai in modo casuale e altri no. Ogni caso rappresenta un diverso partecipante al programma.

Appendice

B

Notices

Patent No. 7,023,453

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: SPSS INC., AN IBM COMPANY, PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. SPSS Inc. may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-SPSS and non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this SPSS Inc. product and use of those Web sites is at your own risk.

When you send information to IBM or SPSS, you grant IBM and SPSS a nonexclusive right to use or distribute the information in any way it believes appropriate without incurring any obligation to you.

Information concerning non-SPSS products was obtained from the suppliers of those products, their published announcements or other publicly available sources. SPSS has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-SPSS products. Questions on the capabilities of non-SPSS products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to SPSS Inc., for the purposes of developing,

using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. SPSS Inc., therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. SPSS Inc. shall not be liable for any damages arising out of your use of the sample programs.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks of IBM Corporation, registered in many jurisdictions worldwide. A current list of IBM trademarks is available on the Web at *http://www.ibm.com/legal/copytrade.shmtl*.

SPSS is a trademark of SPSS Inc., an IBM Company, registered in many jurisdictions worldwide.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Java and all Java-based trademarks and logos are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

This product uses WinWrap Basic, Copyright 1993-2007, Polar Engineering and Consulting, *http://www.winwrap.com*.

Other product and service names might be trademarks of IBM, SPSS, or other companies.

Adobe product screenshot(s) reprinted with permission from Adobe Systems Incorporated.

Microsoft product screenshot(s) reprinted with permission from Microsoft Corporation.

Bibliografia

Bell, E. H. 1961. *Social foundations of human behavior: Introduction to the study of sociology*. New York: Harper & Row.

Blake, C. L., e C. J. Merz. 1998. "UCI Repository of machine learning databases." Available at http://www.ics.uci.edu/~mlearn/MLRepository.html.

Breiman, L., e J. H. Friedman. 1985. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80, .

Collett, D. 2003. *Modelling survival data in medical research*, 2 ed. Boca Raton: Chapman & Hall/CRC.

Green, P. E., e V. Rao. 1972. *Applied multidimensional scaling*. Hinsdale, Ill.: Dryden Press.

Green, P. E., e Y. Wind. 1973. *Multiattribute decisions in marketing: A measurement approach*. Hinsdale, Ill.: Dryden Press.

Guttman, L. 1968. A general nonmetric technique for finding the smallest coordinate space for configurations of points. *Psychometrika*, 33, .

Hartigan, J. A. 1975. *Clustering algorithms*. New York: John Wiley and Sons.

Hastie, T., e R. Tibshirani. 1990. *Generalized additive models*. London: Chapman and Hall.

Kennedy, R., C. Riquier, e B. Sharp. 1996. Practical applications of correspondence analysis to categorical data in market research. *Journal of Targeting, Measurement and Analysis for Marketing*, 5, .

McCullagh, P., e J. A. Nelder. 1989. *Generalized Linear Models*, 2nd ed. London: Chapman & Hall.

Prezzo, R. H., e D. L. Bouffard. 1974. Behavioral appropriateness and situational constraints as dimensions of social behavior. *Journal of Personality and Social Psychology*, 30, .

Rickman, R., N. Mitchell, J. Dingman, e J. E. Dalen. 1974. Changes in serum cholesterol during the Stillman Diet. *Journal of the American Medical Association*, 228, .

Rosenberg, S., e M. P. Kim. 1975. The method of sorting as a data-gathering procedure in multivariate research. *Multivariate Behavioral Research*, 10, .

Van der Ham, T., J. J. Meulman, D. C. Van Strien, e H. Van Engeland. 1997. Empirically based subgrouping of eating disorders in adolescents: A longitudinal perspective. *British Journal of Psychiatry*, 170, .

Verdegaal, R. 1985. *Meer sets analyse voor kwalitatieve gegevens (in Dutch)*. Leiden: Department of Data Theory, University of Leiden.

Ware, J. H., D. W. Dockery, A. Spiro III, F. E. Speizer, e B. G. Ferris Jr.. 1984. Passive smoking, gas cooking, and respiratory health of children living in six cities. *American Review of Respiratory Diseases*, 129, .

Indice