IBM SPSS Decision Trees 19

Note: Before using this information and the product it supports, read the general information under Notices a pag. 114.

Prefazione

IBM® SPSS® Statistics è un sistema completo per l'analisi dei dati. Il modulo aggiuntivo opzionale Decision Trees include le tecniche di analisi aggiuntive descritte nel presente manuale. Il modulo aggiuntivo Decision Trees 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

1 Creazione di Alberi decisionali 1

Selezione delle categorie 6
Convalida 8
Criteri di espansione dell'albero 9
Limiti di crescita 9
Criteri CHAID 10
Criteri CRT 12
Criteri QUEST 14
Taglio degli alberi 15
Surrogati 16
Opzioni 16
Costi classificazione errata 17
Profitti 18
Probabilità a priori 19
Punteggi 21
Valori mancanti 22
Salvataggio delle informazioni del modello 24
Output 25
Visualizzazione dell'albero 25
Statistiche 27
Grafici 31
Regole di selezione e di punteggio 37

2 Editor albero 39

Utilizzo di alberi di grandi dimensioni 40
Mappa albero 41
Scaling della visualizzazione dell'albero 42
Finestra Riepilogo nodi 42
Controllo delle informazioni visualizzate nell'albero 43
Modifica dei colori dell'albero e dei caratteri del testo 44

Regole di selezione e di punteggio dei casi 46
Applicazione di filtri ai casi 46
Salvataggio di regole di selezione e di punteggio 47

Parte II: Esempi

3 Ipotesi sui dati e requisiti 50

Effetti del livello di misurazione sui modelli di alberi 50
Assegnazione permanente del livello di misurazione 53
Variabili con livello di misurazione sconosciuto 54
Effetti delle etichette dei valori sui modelli di alberi 54
Assegnazione di etichette dei valori a tutti i valori 56

4 Utilizzo degli alberi decisionali per la valutazione del rischio di credito 58

Creazione del modello 58
Creazione del modello di albero CHAID 58
Selezione delle categorie obbiettivo 59
Specificazione dei criteri di espansione dell'albero 60
Selezione di output aggiuntivo 61
Salvataggio di valori attesi 63
Valutazione del modello 64
Tabella Riepilogo del modello 65
Diagramma ad albero 66
Tabella albero 67
Guadagni per i nodi 69
Grafico Guadagni 70
Grafico indice 71
Stima del rischio e classificazione 72
Valori attesi 73
Perfezionamento del modello 74
Selezione di casi nei nodi 74
Esame dei casi selezionati 75
Assegnazione dei costi ai risultati 78
Riepilogo 82

5 Creazione di un modello di credito 83

Creazione del modello 83
Valutazione del modello 85
Riepilogo del modello. 86
Diagramma del modello di albero. 87
Stima del rischio 88
Applicazione del modello a un altro file di dati. 89
Riepilogo 92

6 Valori mancanti nei modelli di albero 93

Valori mancanti con CHAID. 94
Risultati CHAID 96
Valori mancanti con CRT. 97
Risultati CRT 100
Riepilogo 102

Appendici

A File di esempio 103

B Notices 114

Indice 116

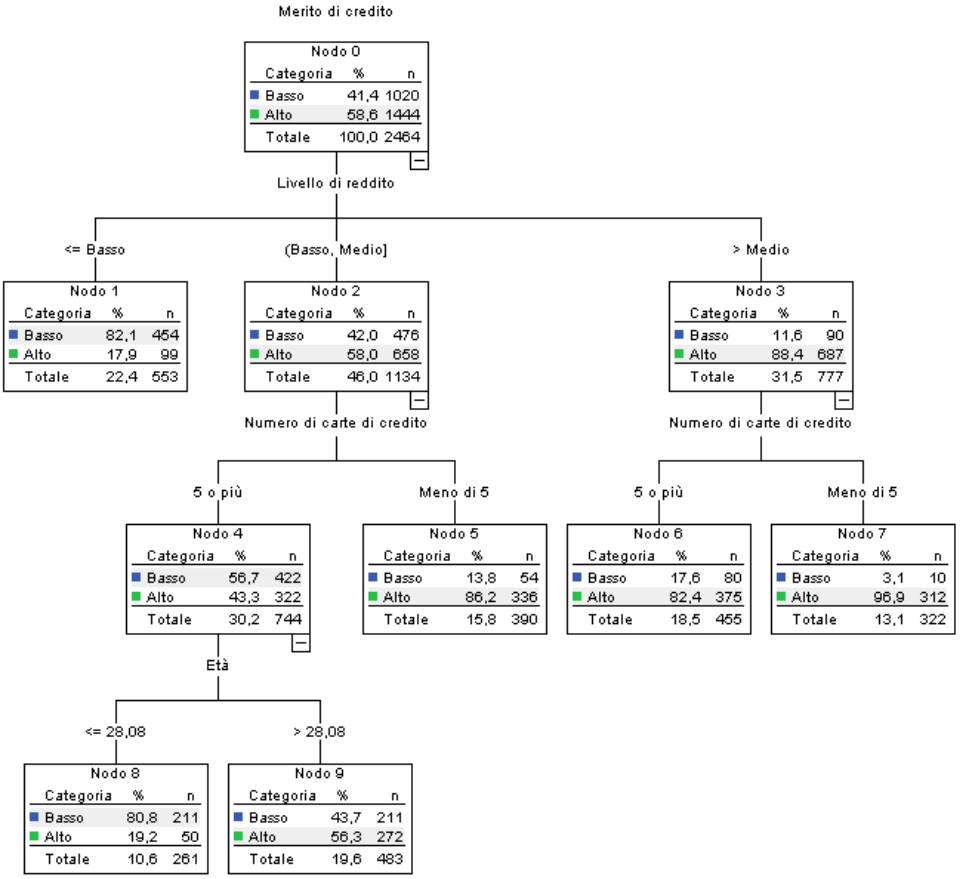
Parte I:
Manuale dell'utente

Capitolo

1

Creazione di Alberi decisionali

Figura 1-1
Albero decisionale



La procedura Albero decisionale crea un modello di classificazione basato su alberi. Classifica i casi in gruppi o prevede i valori di una variabile dipendente (di destinazione) in base ai valori di variabili (predittore) indipendenti. La procedura offre strumenti di validazione per l'analisi di classificazione confermativa ed esplorativa.

È possibile utilizzare la procedura per eseguire le seguenti operazioni:

Segmentazione. Identifica gli individui che appartengono a un determinato gruppo.

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

Previsione. Crea regole e le utilizza per prevedere eventi futuri, ad esempio la probabilità che qualcuno non rimborsi un prestito o il valore di rivendita potenziale di un veicolo o di un'immobile.

Riduzione dei dati ed esame delle variabili. Seleziona un sottoinsieme utile di predittori per un insieme ampio di variabili da utilizzare nella creazione di un modello parametrico formale.

Identificazione delle interazioni. Identifica le relazioni pertinenti solo a determinati sottogruppi e le specifica in un modello parametrico formale.

Unione delle categorie e discretizzazione delle variabili continue. Ricodifica le categorie di predittori e le variabili continue con una perdita minima di informazioni.

Esempio. Una banca desidera categorizzare i richiedenti di credito in base al fatto che rappresentino o meno un rischio di credito ragionevole. In base a vari fattori, comprese le valutazioni di credito note di clienti precedenti, è possibile creare un modello per prevedere se è probabile che i clienti futuri non rimborsino i propri prestiti.

Un'analisi basata su diagrammi ad albero offre alcune funzioni interessanti:

- consente di identificare gruppi omogenei a basso o ad alto rischio.
- Semplifica la creazione di regole per l'esecuzione di previsioni relative a singoli casi.

Considerazioni sui dati

Dati. Le variabili dipendenti ed indipendenti possono essere:

- **Nominale.** Una variabile può essere considerata nominale quando i relativi valori rappresentano categorie prive di ordinamento intrinseco, per esempio l'ufficio di una società, Tra gli esempi di variabili nominali troviamo la regione, il codice postale e la religione.
- **Ordinale.** Una variabile può essere considerata ordinale quando i relativi valori rappresentano categorie con qualche ordinamento intrinseco, per esempio i gradi di soddisfazione per un servizio, da molto insoddisfatto a molto soddisfatto, i punteggi di atteggiamento corrispondenti a gradi di soddisfazione o fiducia e i punteggi di preferenza.
- **Scala.** Una variabile può essere considerata di scala (continua) quando i relativi valori rappresentano categorie ordinate con una metrica significativa, tale che i confronti fra le distanze dei relativi valori siano appropriati. Esempi di variabili di scala sono l'età espressa in anni o il reddito espresso in migliaia di Euro.

Ponderazione Se la ponderazione è attiva, i pesi frazionari vengono arrotondati all'intero più vicino; di conseguenza, ai casi con un peso inferiore a 0,5 viene assegnato un peso pari a 0 e di conseguenza vengono esclusi dall'analisi.

Assunzioni. La procedura presuppone che il livello di misurazione appropriato sia stato assegnato a tutte le variabili dell'analisi; alcune funzioni presuppongono che tutti i valori della variabile dipendente inclusi nell'analisi abbiano etichette dei valori definite.

- **Livello di misurazione.** Il livello di misurazione influenza i calcoli dell'albero; di conseguenza a tutte le variabili deve essere assegnato il livello di misurazione appropriato. Per impostazione predefinita, si suppone che le variabili numeriche siano di scala e le variabili stringa nominali,

il che potrebbe non riflettere con precisione il livello di misurazione effettivo. L'icona accanto a ciascuna variabile nell'elenco delle variabili ne identifica il tipo.

 Scala

 Nominale

 Ordinale

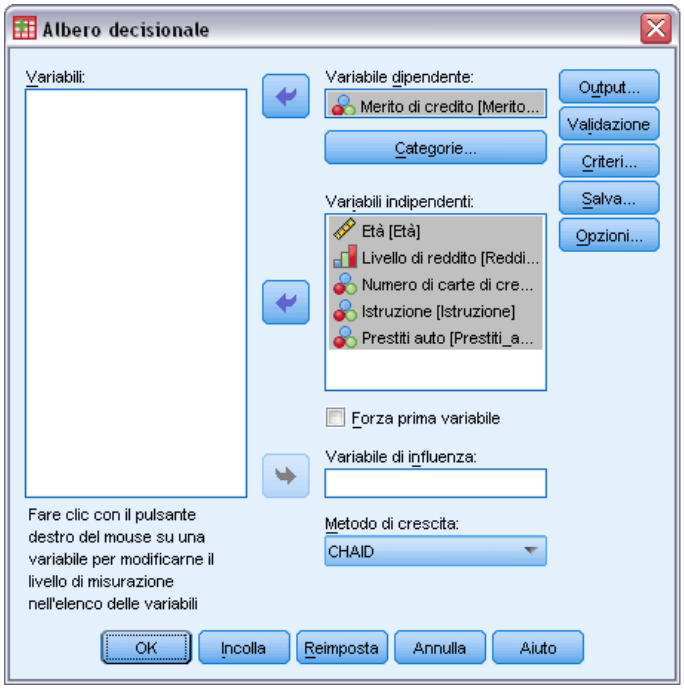
Per modificare temporaneamente il livello di misurazione di una variabile è possibile fare clic con il pulsante destro del mouse sulla variabile nell'elenco di variabili sorgenti e scegliere un livello di misurazione dal menu di scelta rapida.

- **Etichette dei valori.** L'interfaccia della finestra di dialogo per la procedura presuppone che per tutti o per nessuno dei valori non mancanti di una variabile dipendente categoriale (nominale, ordinale) siano state definite etichette dei valori. Alcune funzioni sono disponibili solo se almeno due valori non mancanti della variabile dipendente categoriale dispongono di etichette dei valori. Se per almeno due valori non mancanti sono state definite etichette dei valori, qualsiasi caso con altri valori privi di etichette sarà escluso dall'analisi.

Per ottenere gli alberi decisionali

► Dai menu, scegliere:
Analizza > Classifica > Albero...

Figura 1-2
Finestra di dialogo Albero decisionale



- Selezionare una variabile dipendente.
- Selezionare una o più variabili indipendenti.
- Selezionare un metodo di espansione.

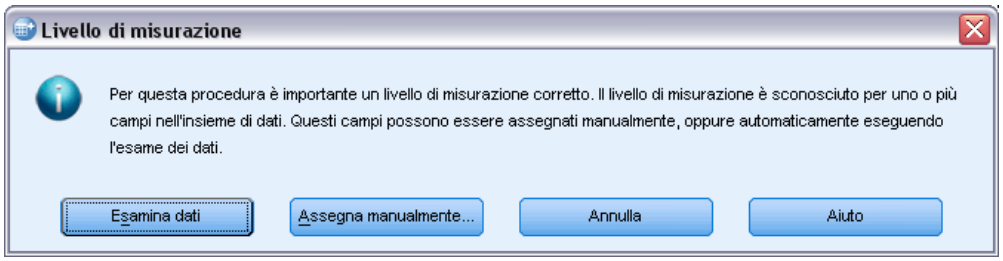
Se lo si desidera, è possibile:

- Modificare il livello di misurazione per qualsiasi variabile nell'elenco sorgente.
- Forzare la prima variabile nell'elenco delle variabili indipendenti nel modello come prima variabile di distinzione.
- Selezionare una variabile di influenza che definisce l'influenza di un caso sul processo di espansione dell'albero. I casi con valori di influenza minori hanno minore influenza, e viceversa. I valori delle variabili di influenza devono essere positivi.
- Convalidare l'albero.
- Personalizzare i criteri di espansione dell'albero.
- Selezionare i numeri dei nodi terminali, i valori attesi e le probabilità previste come variabili.
- Salvare il modello in formato XML (PMML).

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 1-3
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.

Modifica del livello di misurazione

▶ Fare clic con il pulsante destro del mouse sulla variabile nell'elenco sorgente.

▶ Scegliere un livello di misurazione dal menu di scelta rapida popup.

Questa operazione modifica temporaneamente il livello di misurazione per utilizzarlo nella procedura Albero decisionale.

Metodi di espansione

I metodi di espansione disponibili sono:

CHAID. Acronimo di Chi-squared Automatic Interaction Detection. Per ogni passaggio, CHAID scegliere la variabile (predittore) indipendente con la più forte interazione con la variabile dipendente. Le categorie di ogni predittore sono unite se non sono diverse in modo rilevante dalla variabile dipendente.

CHAID esaustivo. Una variante di CHAID che esamina tutte le suddivisioni possibili per ciascun predittore.

CRT. Alberi decisionali e di regressione. CRT divide i dati in segmenti che sono il più possibile omogenei rispetto alla variabile dipendente. Un nodo terminale in cui tutti i casi hanno lo stesso valore per la variabile dipendente è un nodo omogeneo o "puro".

QUEST. Acronimo di Quick, Unbiased, Efficient Statistical Tree. Metodo che esegue i calcoli molto velocemente ed evita la polarizzazione degli altri metodi a favore dei predittori con molte categorie È possibile specificarlo solo se la variabile numerica è nominale.

Ogni metodo presenta vantaggi e limitazioni, tra i quali:

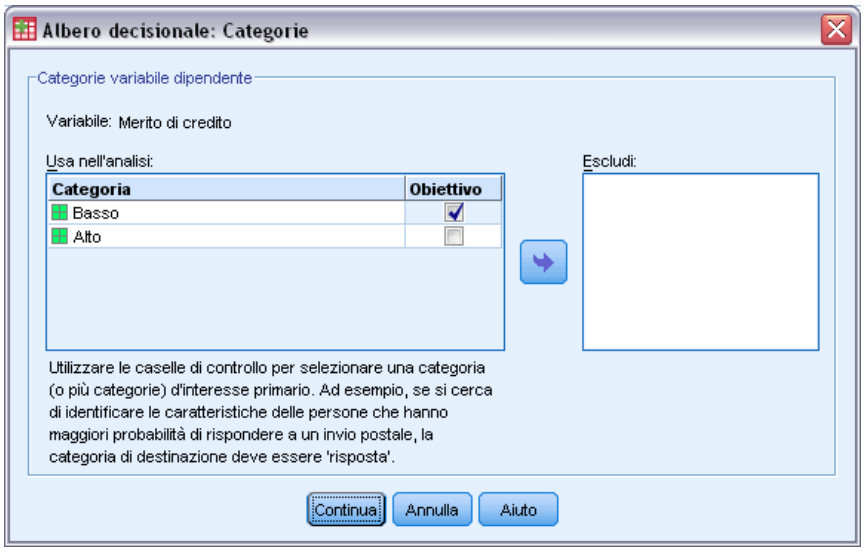
	CHAID*	CRT	QUEST
Basato su chi-quadrato**	S		
Variabili (predittori) indipendenti di surrogati		S	S
Taglio degli alberi		S	S
Divisione dei nodi a più vie	S		
Divisione dei nodi binaria		S	S
Variabili di influenza	S	S	
Probabilità a priori		S	S
Costi di errata classificazione	S	S	S
Calcolo rapido	S		S

* Include CHAID esaustivo.

**QUEST utilizza inoltre una misura di chi-quadrato per le variabili indipendenti nominali.

Selezione delle categorie

Figura 1-4
Finestra di dialogo Categorie



Per variabili dipendenti (nominali, ordinali) categoriali, è possibile:

- Controllare quali categorie sono incluse nell'analisi.
- Identificare le categorie obbiettivo di interesse.

Includere/escludere categorie

È possibile limitare l'analisi a categorie specifiche della variabile dipendente.

- I casi con valori della variabile dipendente nell'elenco Escludi non vengono inclusi nell'analisi.
- Per variabili dipendenti nominali, è possibile inoltre includere le categorie mancanti definite dall'utente nell'analisi. Per impostazione predefinita, le categorie mancanti definite dall'utente vengono visualizzate nell'elenco Escludi.

Categorie obbiettivo

Se l'opzione è selezionata, le categorie vengono considerate come categorie di interesse principale nell'analisi. Ad esempio, se si è interessati principalmente all'identificazione delle persone che più probabilmente non rimborseranno un prestito, selezionare la categoria di valutazione creditizia "negativa" come categoria obbiettivo.

- Non esiste una categoria obbiettivo predefinita. Se non è selezionata alcuna categoria, alcune opzioni relative alle regole di classificazione e alcuni output correlati ai guadagni non sono disponibili.
- Se sono selezionate più categorie, vengono prodotte tabelle di guadagno e grafici distinti per ciascuna categoria obbiettivo.
- La designazione di una o più categorie come categorie obbiettivo non ha alcun effetto sul modello dell'albero, sulla stima del rischio o sui risultati di errata classificazione.

Categorie ed etichette dei valori

La finestra di dialogo richiede etichette dei valori definite per la variabile dipendente. Non è disponibile a meno che per almeno due valori della variabile dipendente categoriale siano state definite etichette dei valori.

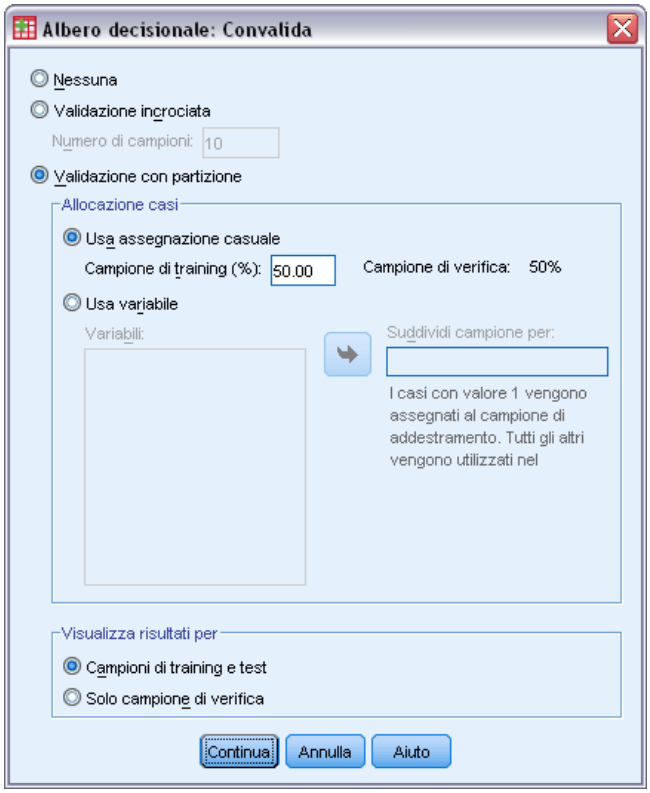
Per includere/escludere categorie e selezionare categorie obbiettivo

▶ Nella finestra di dialogo principale Albero decisionale, selezionare una variabile dipendente (nominale, ordinale) categoriale con due o più etichette dei valori definite.

▶ Fare clic su Categorie.

Convalida

Figura 1-5
Finestra di dialogo Convalida



La convalida consente di valutare in che modo la struttura ad albero generalizza i dati in riferimento a una popolazione più ampia. I metodi di convalida disponibili sono due: convalida incrociata e convalida con suddivisione.

Validazione incrociata

La convalida incrociata divide il campione in vari sottocampioni, o **campioni**. I modelli ad albero vengono quindi generati escludendo di volta in volta i dati da ciascun sottocampione. Il primo albero si basa su tutti i casi eccetto quelli contenuti nel primo campione, il secondo albero si basa su tutti i casi eccetto quelli contenuti nel secondo campione e così via. Il rischio di errata classificazione per ciascun albero viene stimato applicando l'albero al sottocampione escluso al momento della generazione dell'albero.

- È possibile specificare un numero massimo di 25 campioni. Maggiore è il valore, minore il numero di casi esclusi per ciascun modello di albero.
- La convalida incrociata genera un unico modello di albero finale. La stima del rischio sulla convalida incrociata per l'albero finale è calcolata come la media dei rischi per tutti gli alberi.

Convalida con suddivisione

La convalida con suddivisione determina la generazione del modello utilizzando un campione di addestramento e la sua verifica su un campione di controllo.

- È possibile specificare la dimensione del campione di addestramento, espressa come percentuale della dimensione totale del campione, o una variabile che divida il campione in campioni di addestramento e di verifica.
- Se si utilizza una variabile per definire i campioni di addestramento e di verifica, i casi con valore 1 per la variabile vengono assegnati al campione di addestramento, mentre tutti gli altri vengono assegnati al campione di verifica. La variabile non può essere la variabile dipendente, la variabile peso, la variabile di influenza o una variabile indipendente forzata.
- È possibile visualizzare i risultati per i campioni di addestramento e di verifica o solo per questi ultimi.
- La convalida con suddivisione deve essere utilizzata con attenzione su file dati di piccole dimensioni (con un numero ridotto di casi). Dimensioni ridotte dei campioni di addestramento possono generare modelli di scarsa qualità, poiché il numero di casi in alcune categorie potrebbe non essere sufficiente a un'espansione adeguata dell'albero.

Criteri di espansione dell'albero

I criteri di espansione disponibili possono variare in base al metodo di espansione, al livello di misurazione della variabile dipendente o a una combinazione dei due elementi.

Limiti di crescita

Figura 1-6
Finestra di dialogo Criteri, scheda Limiti di crescita

La scheda Limiti di crescita consente di limitare il numero dei livelli dell'albero e di controllare il numero minimo di casi per i nodi genitore e figlio.

Massima profondità struttura (livelli). Controlla il numero massimo di livelli di espansione al di sotto del nodo radice. L'impostazione Automatico limita l'albero a tre livelli sotto il nodo radice per i metodi CHAID e CHAID esaustivo e a cinque livelli per i metodi CRT e QUEST.

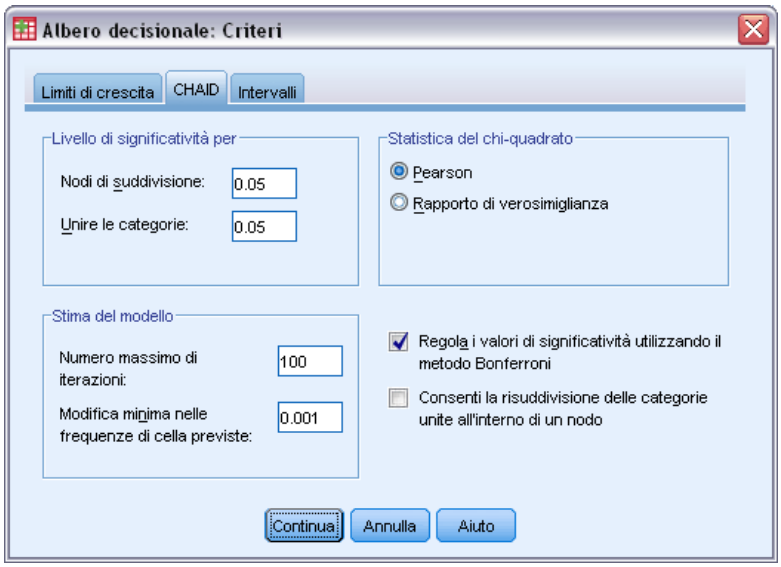
Numero minimo di casi. Controlla il numero minimo di casi per i nodi. I nodi che non rispondono a questi criteri non vengono divisi.

- L'aumento dei valori minimi tende a generare alberi con un numero inferiore di nodi.
- La riduzione dei valori minimi tende a generare alberi con un numero superiore di nodi.

Per i file dati con un numero ridotto di casi, i valori predefiniti di 100 casi per i nodi genitore e di 50 casi per i nodi figlio possono generare alberi senza nodi al di sotto del nodo radice; in questo caso, riducendo i valori minimi si possono ottenere risultati più significativi.

Criteri CHAID

Figura 1-7
Finestra di dialogo Criteri, scheda CHAID



Per i metodi CHAID e CHAID esaustivo, è possibile controllare:

Livello di significatività. È possibile controllare il valore di significatività per la divisione dei nodi e l'unione delle categorie. Per entrambi i criteri, il livello di significatività predefinito è 0,05.

- Per la divisione dei nodi, il valore deve essere maggiore di 0 e minore di 1. Valori inferiori tendono a generare alberi con un numero inferiore di nodi.
- Per l'unione delle categorie, il valore deve essere maggiore di 0 e minore o uguale a 1. Per impedire l'unione delle categorie, specificare il valore 1. Per una variabile indipendente di scala, questo significa che il numero di categorie per la variabile nell'albero finale è il numero specificato di intervalli (il numero predefinito è 10). Per ulteriori informazioni, vedere l'argomento Intervalli di scala per l'analisi CHAID a pag. 11.

Statistica chi-quadrato. Per le variabili dipendenti ordinali, il chi-quadrato per la determinazione della divisione dei nodi e l'unione delle categorie viene calcolato utilizzando il metodo del rapporto di verosimiglianza. Per variabili dipendenti nominali è possibile selezionare il metodo:

- **Pearson** Questo metodo offre calcoli più rapidi ma deve essere utilizzato con attenzione su campioni di dimensioni ridotte. È il metodo predefinito.
- **Rapporto di verosimiglianza.** È un metodo più solido del precedente, ma richiede più tempo per i calcoli. È il metodo di elezione per campioni di piccole dimensioni.

Stima del modello. Per variabili dipendenti nominali e ordinali è possibile specificare:

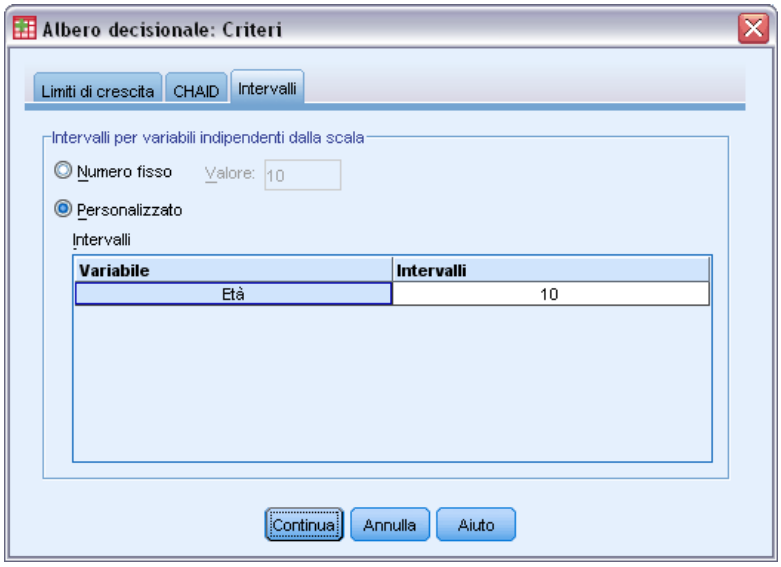
- **Numero massimo di iterazioni.** Il valore predefinito è 100. Se l'espansione dell'albero si arresta a causa del raggiungimento del numero massimo di iterazioni, potrebbe essere consigliabile aumentare tale valore o modificare uno o più tra gli altri criteri che controllano l'espansione dell'albero.
- **Modifica minima nelle frequenze attese di cella.** Il valore deve essere maggiore di 0 e minore di 1. Il valore predefinito è 0,05. Valori minori tendono a generare alberi con un numero inferiore di nodi.

Correzione dei valori di significatività utilizzando il metodo di Bonferroni. Per confronti multipli, i valori di significatività per i criteri di unione e di divisione vengono corretti tramite il metodo di Bonferroni. È l'impostazione di default.

Consenti la ridivisione delle categorie unite all'interno di un nodo. Salvo l'unione delle categorie venga impedita esplicitamente, la procedura tenterà di unire le categorie di variabili (predittore) indipendenti per generare l'albero più semplice descrittivo del modello. L'opzione consente alla procedura di ridividere le categorie unite se questo offre una soluzione migliore.

Intervalli di scala per l'analisi CHAID

Figura 1-8
Finestra di dialogo Criteri, scheda Intervalli



Nell'analisi CHAID, le variabili (predittore) indipendenti vengono sempre segmentate in gruppi discreti (ad esempio 0–10, 11–20, 21–30 e così via) prima dell'analisi. È possibile controllare il numero iniziale/massimo dei gruppi (sebbene la procedura possa unire gruppi consecutivi dopo la divisione iniziale):

- **Numero fisso.** Tutte le variabili indipendenti di scala vengono inizialmente segmentate nello stesso numero di gruppi. Il valore di default è 10.
- **Personalizzata.** Ciascuna variabile indipendente di scala viene inizialmente segmentata nel numero di gruppi specificato per la variabile.

Per specificare intervalli per variabili indipendenti di scala

▶ Nella finestra di dialogo principale Albero decisionale selezionare una o più variabili indipendenti di scala.

▶ Come metodo di espansione scegliere CHAID o CHAID esaustivo.

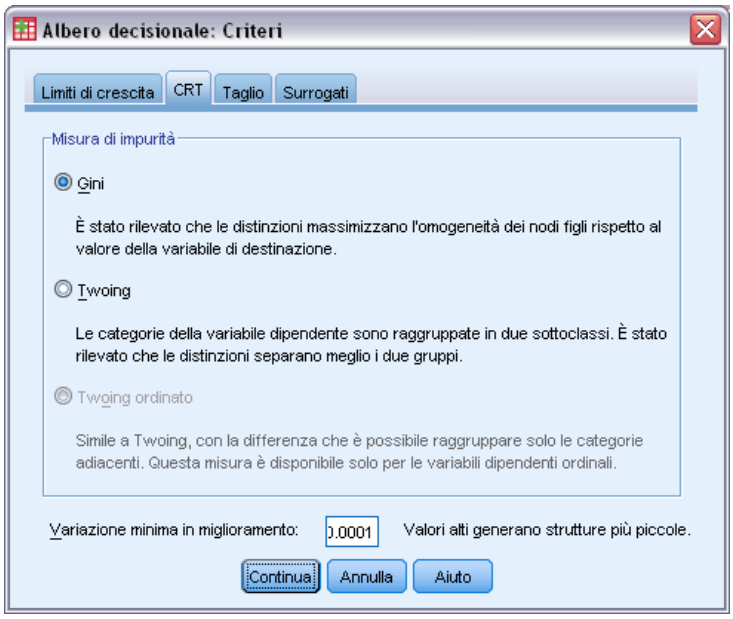
▶ Fare clic su Criteri.

▶ Fare clic sulla scheda Intervalli.

Nell'analisi CRT e QUEST, tutte le divisioni sono binarie e le variabili indipendenti ordinali e di scala vengono gestite nello stesso modo; di conseguenza, non è possibile specificare un numero di intervalli per le variabili indipendenti di scala.

Criteri CRT

Figura 1-9
Finestra di dialogo Criteri, scheda CRT



Il metodo di espansione CRT tenta di massimizzare l'omogeneità all'interno del nodo. La misura in cui un nodo non rappresenta un sottoinsieme omogeneo di casi è un indicatore di **impurità**. Ad esempio, un nodo terminale in cui tutti i casi hanno lo stesso valore per la variabile dipendente è un nodo omogeneo che non richiede divisioni ulteriori, in quanto "puro".

È possibile selezionare il metodo utilizzato per misurare l'impurità e la riduzione minima nell'impurità richiesta per la divisione dei nodi.

Misura dell'impurità Per variabili dipendenti di scala, viene utilizzata la misura di impurità Least-Squared Deviation (LSD). Viene calcolato allo stesso modo della varianza all'interno del nodo, adeguata in base alla ponderazione o ai valori di influenza.

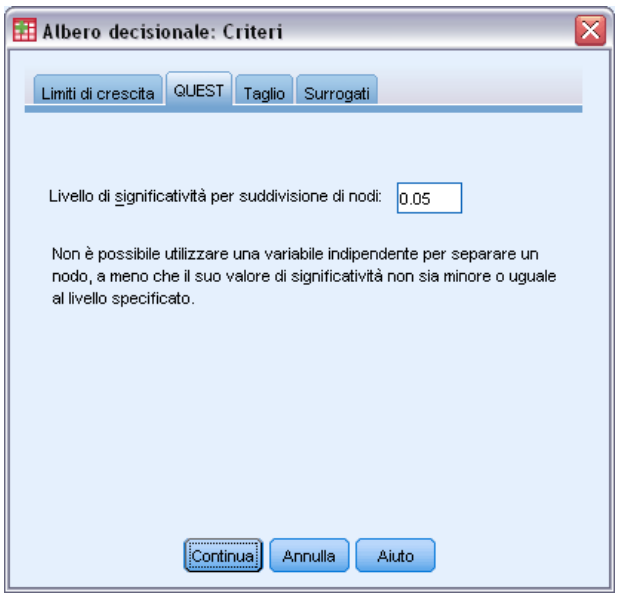
Per variabili dipendenti (nominali, ordinali) categoriali, è possibile selezionare la misura dell'impurità:

- **Gini** Vengono individuate le divisioni che massimizzano l'omogeneità dei nodi figlio rispetto al valore della variabile dipendente. Il metodo Gini si basa sulle probabilità quadratiche di appartenenza per ciascuna categoria della variabile dipendente. Questo valore raggiunge il minimo (zero) quando tutti i casi di un nodo rientrano in un'unica categoria. È la misura predefinita.
- **Twoing**. Le categorie della variabile dipendente sono raggruppate in due sottoclassi. Vengono individuate le divisioni migliori tra i due gruppi.
- **Twoing ordinato**. Analogo al Twoing, fatta eccezione per il fatto che possono essere raggruppate solo categorie adiacenti. La misura è disponibile solo per le variabili dipendenti ordinali.

Modifica minima nel miglioramento. La riduzione minima nell'impurità richiesta per la divisione di un nodo. Il valore di default è 0.0001. Valori maggiori tendono a generare alberi con un numero inferiore di nodi.

Criteri QUEST

Figura 1-10
Finestra di dialogo Criteri, scheda QUEST



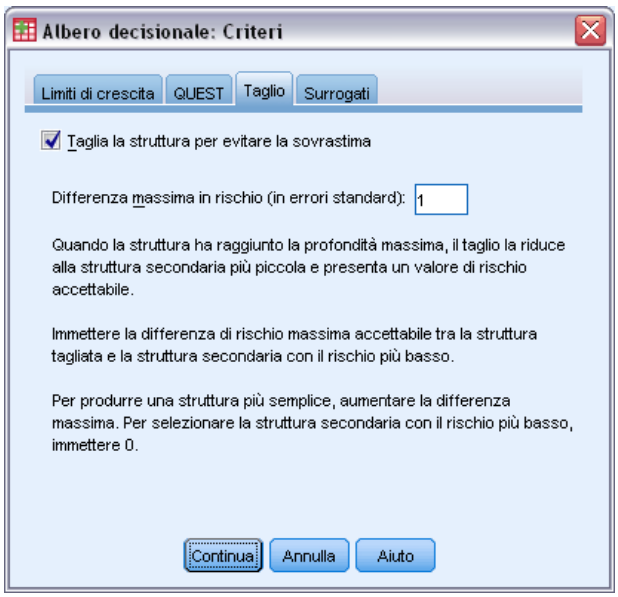
Per il metodo QUEST è possibile specificare il livello di significatività per la divisione dei nodi. Non è possibile utilizzare una variabile indipendente per la divisione dei nodi a meno che il livello di significatività non sia minore o uguale al valore specificato. Il valore deve essere maggiore di 0 e minore di 1. Il valore predefinito è 0,05. Valori inferiori tenderanno a escludere un maggior numero di variabili indipendenti dal modello finale.

Per specificare i criteri QUEST

- Nella finestra di dialogo principale Albero decisionale selezionare una variabile dipendente nominale.
- Come metodo di espansione scegliere QUEST.
- Fare clic su Criteri.
- Fare clic sulla scheda QUEST.

Taglio degli alberi

Figura 1-11
Finestra di dialogo Criteri, scheda Taglio



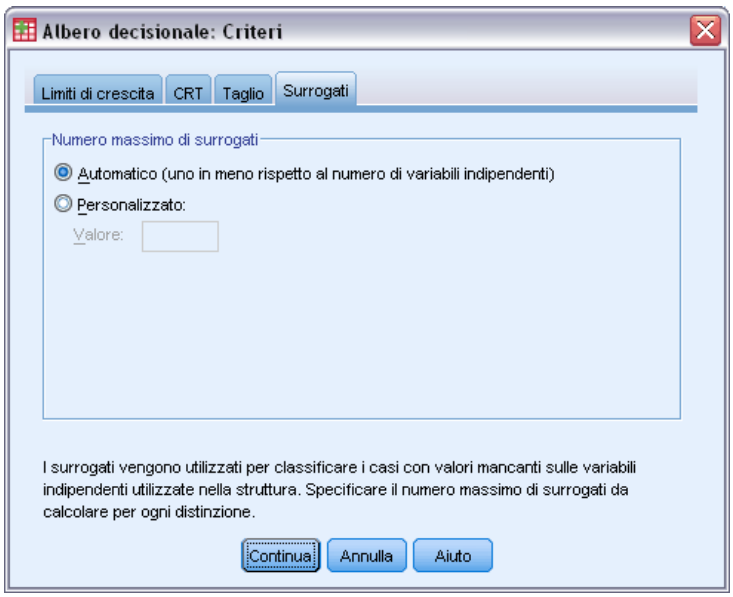
Con i metodi CRT e QUEST è possibile evitare di sovraccaricare il modello **tagliando** l'albero: l'albero si espande fino al raggiungimento dei criteri di arresto, dopodiché il taglio viene eseguito automaticamente in base al sottoalbero più piccolo in base alla differenza massima specificata nel rischio. Il valore di rischio è espresso in errori standard. Il valore predefinito è 1. Deve essere non negativo. Per ottenere un sottoalbero con rischio minimo, specificare 0.

Confronto tra tagliare e nascondere i nodi

Quando si crea un albero tagliato, eventuali nodi tagliati non saranno disponibili nell'albero finale. È possibile nascondere e mostrare in modo interattivo i nodi figlio selezionati nell'albero finale, ma non mostrare i nodi tagliati durante il processo di creazione dell'albero. Per ulteriori informazioni, vedere l'argomento Editor albero in il capitolo 2 a pag. 39.

Surrogati

Figura 1-12
Finestra di dialogo Criteri, scheda Surrogati



CRT e QUEST possono utilizzare i **surrogati** per le variabili (predittore) indipendenti. Per in casi in cui il valore per la variabile è mancante, per la classificazione sono utilizzate altre variabili indipendenti con associazioni ++elevate con la variabile originale. Questi predittori alternativi sono detti surrogati. È possibile specificare il numero massimo di surrogati da utilizzare nel modello.

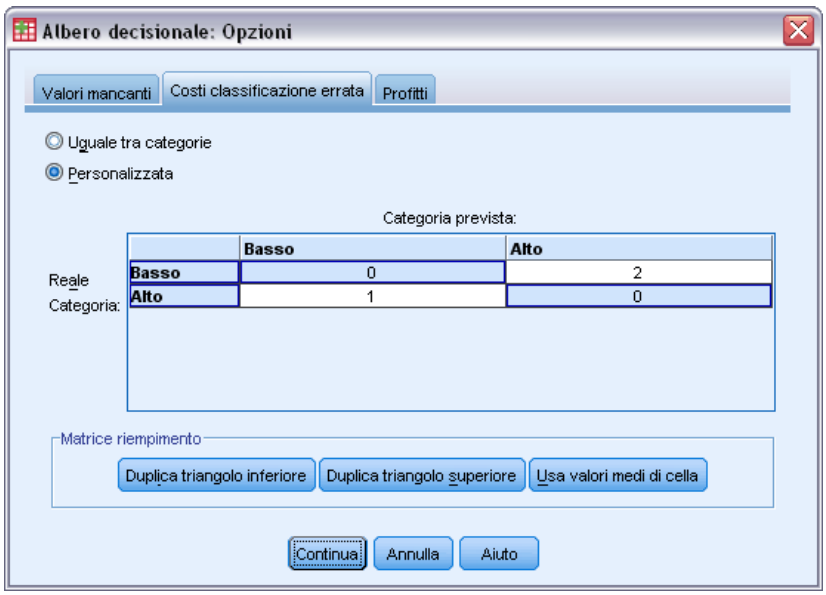
- Per impostazione predefinita, il numero massimo di surrogati è pari al numero di variabili indipendenti meno uno. In altre parole, per ciascuna variabile indipendente, tutte le altre possono essere utilizzate come surrogati.
- Se non si desidera utilizzare surrogati nel modello, specificare 0 come numero di surrogati.

Opzioni

Le opzioni disponibili possono variare in base al metodo di espansione, al livello di misurazione della variabile dipendente e/o all'esistenza di etichette dei valori definite per i valori della variabile dipendente.

Costi classificazione errata

Figura 1-13
Finestra di dialogo Opzioni, scheda Costi di errata classificazione



Per variabili dipendenti (nominali, ordinali) categoriali, i costi di errata classificazione consentono di includere informazioni sulla penalità associata alla classificazione errata. Ad esempio:

- Il costo di negare il credito a un cliente meritevole sarà probabilmente diverso dal costo di concedere il credito a un cliente che si rivelerà inadempiente.
- Il costo dovuto all'errata classificazione di un singolo ad alto rischio di malattia cardiaca come a basso rischio è molto maggiore del costo dovuto all'errata classificazione di individui a basso rischio come ad alto rischio.
- Il costo di inviare un mailing di massa a qualcuno che probabilmente non risponderà sarà normalmente piuttosto basso, mentre il costo del mancato invio della stessa comunicazione a qualcuno che probabilmente avrebbe risposto è relativamente maggiore, in termini di mancato profitto.

Costi di errata classificazione ed etichette dei valori

La finestra di dialogo non è disponibile a meno che per almeno due valori della variabile dipendente categoriale siano state definite etichette dei valori.

Per specificare i costi di errata classificazione

▶ Nella finestra di dialogo principale Albero decisionale, selezionare una variabile dipendente (nominale, ordinale) categoriale con due o più etichette dei valori definite.

▶ Fare clic su Opzioni.

▶ Fare clic sulla scheda Costi errata classificazione.

▶ Fare clic su Personalizzato.

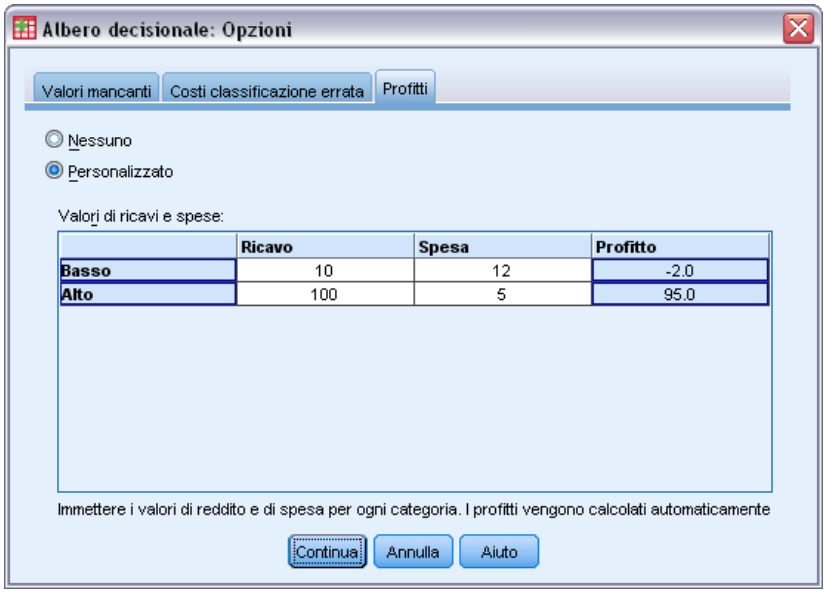
▶ Inserire uno o più costi di errata classificazione nella griglia. I valori devono essere non negativi (le classificazione corrette, rappresentate sulla diagonale, sono sempre 0).

Riempimento matrice. In molti casi, è necessario che i costi siano simmetrici—ossia che il costo dovuto all'errata classificazione di A come B corrisponda al costo dovuto all'errata classificazione di B come A. I seguenti comandi facilitano la selezione di una matrice di costi simmetrici:

- **Duplica triangolo inferiore.** Copia i valori del triangolo inferiore della matrice (sotto la diagonale) nelle corrispondenti celle triangolari superiori.
- **Duplica triangolo superiore.** Copia i valori del triangolo superiore della matrice (sopra la diagonale) nelle corrispondenti celle triangolari inferiori.
- **Utilizza valori di cella medi.** Per ogni cella di ciascuna metà della matrice, viene eseguita la media tra i due valori (triangolare superiore e inferiore) e tale media sostituisce entrambi i valori. Ad esempio, se il costo dovuto all'errata classificazione di A come B equivale a 1 e il costo dovuto all'errata classificazione di B come A equivale a 3, il comando sostituirà entrambi i valori con la media (1+3)/2 = 2.

Profitti

Figura 1-14
Finestra di dialogo Opzioni, scheda Profitti



Per le variabili dipendenti categoriali, è possibile assegnare i valori per i ricavi e le spese ai livelli della variabile dipendente.

- Il profitto corrisponde ai ricavi meno le spese.
- I valori relativi al profitto influenzano i valori relativi il profitto medio e il ROI (return on investment) nelle tabelle dei guadagni. Non influenzano nemmeno la struttura del modello dell'albero di base.
- I valori relativi a ricavi e spese devono essere numerici ed essere specificati per tutte le categorie della variabile dipendente visualizzate nella griglia.

Profitti ed etichette dei valori

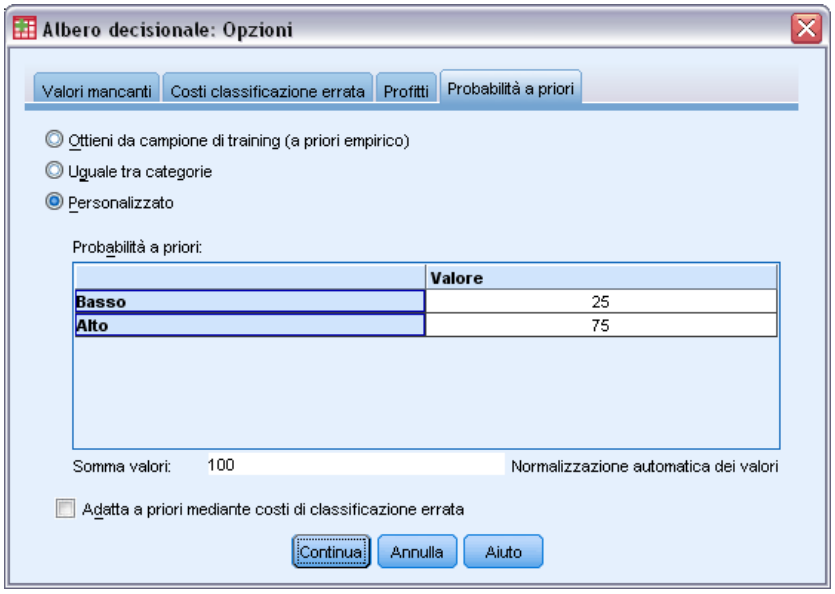
La finestra di dialogo richiede etichette dei valori definite per la variabile dipendente. Non è disponibile a meno che per almeno due valori della variabile dipendente categoriale siano state definite etichette dei valori.

Per specificare i profitti

- Nella finestra di dialogo principale Albero decisionale, selezionare una variabile dipendente (nominale, ordinale) categoriale con due o più etichette dei valori definite.
- Fare clic su Opzioni.
- Fare clic sulla scheda Profitti.
- Fare clic su Personalizzato.
- Inserire i valori relativi a ricavi e spese per tutte le categorie di variabili dipendenti elencate nella griglia.

Probabilità a priori

Figura 1-15
Finestra di dialogo Opzioni, scheda Probabilità a priori



Per gli alberi CRT e QUEST con variabili dipendenti categoriali, è possibile specificare le probabilità a priori di appartenenza al gruppo. Le **probabilità a priori** sono stime della frequenza relativa globale per ciascuna categoria della variabile dipendente prima di conoscere qualsiasi informazione sui valori delle variabili (predittori) indipendenti. L'utilizzo delle probabilità a priori agevola la correzione di un'espansione dell'albero causata da dati del campione non rappresentativi dell'intera popolazione.

Ottieni da campione di addestramento (a priori empirico). Utilizzare questa impostazione se la distribuzione dei valori delle variabili dipendenti nel file dati è rappresentativa della distribuzione della popolazione. Se si utilizza la convalida con suddivisione, viene utilizzata la distribuzione dei casi nel campione di addestramento.

Nota: Poiché nella convalida con suddivisione i casi vengono assegnati in modo casuale al campione di addestramento, la distribuzione effettiva dei casi nel campione non sarà nota in anticipo. Per ulteriori informazioni, vedere l'argomento Convalida a pag. 8.

Uguale per tutte le categorie. Utilizzare questa impostazione se le categorie della variabile dipendente sono rappresentate in modo uguale nella popolazione. Ad esempio, in presenza di quattro categorie, circa il 25% dei casi appartengono a ciascuna categoria.

Personalizzata. Inserire un valore non negativo per ciascuna categoria della variabile dipendente elencata nella griglia. I valori possono essere proprietà, percentuali, conteggi di frequenze o qualsiasi altro valore che rappresenti la distribuzione dei valori tra le categorie.

Adegua le probabilità a priori utilizzando i costi di errata classificazione. Se si definiscono costi di errata classificazione personalizzati, è possibile adeguare le probabilità a priori in base ai costi stessi. Per ulteriori informazioni, vedere l'argomento Costi classificazione errata a pag. 17.

Profitti ed etichette dei valori

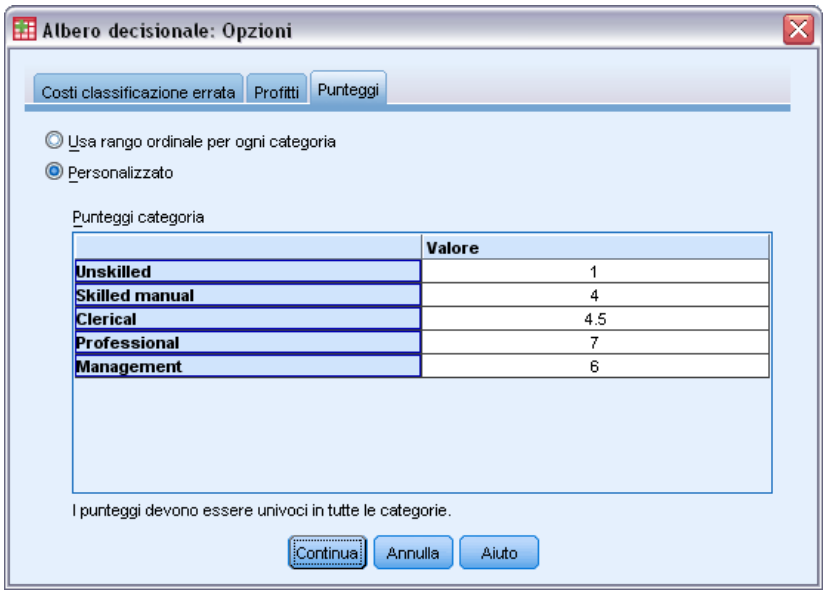
La finestra di dialogo richiede etichette dei valori definite per la variabile dipendente. Non è disponibile a meno che per almeno due valori della variabile dipendente categoriale siano state definite etichette dei valori.

Per specificare le probabilità a priori

- Nella finestra di dialogo principale Albero decisionale, selezionare una variabile dipendente (nominale, ordinale) categoriale con due o più etichette dei valori definite.
- Come metodo di espansione scegliere CRT o QUEST.
- Fare clic su Opzioni.
- Fare clic sulla scheda Probabilità a priori.

Punteggi

Figura 1-16
Finestra di dialogo Opzioni, scheda Punteggi



Per CHAID e CHAID esaustivo con una variabile dipendente ordinale, è possibile assegnare punteggi personalizzati a ciascuna categoria della variabile dipendente. I punteggi definiscono l'ordine e la distanza tra le categorie della variabile dipendente. È possibile utilizzare i punteggi per aumentare o ridurre la distanza relativa tra i valori ordinali o per modificare l'ordine dei valori.

- **Utilizza rango ordinale per ciascuna categoria** Alla categoria più bassa della variabile dipendente viene assegnato il punteggio 1, alla categoria più alta successiva 2 e così via. È l'impostazione di default.
- **Personalizzata.** Inserire un valore di punteggio numerico per ciascuna categoria della variabile dipendente elencata nella griglia.

Esempio

Etichetta valori	Valore originale	Punteggio
Non specializzato	1	1
Operaio specializzato	2	4
Impiegato	3	4.5
Professional	4	7
Dirigenza	5	6

- I punteggi aumentano la distanza relativa tra *Non specializzato* e *Operaio specializzato* e riducono la distanza relativa tra *Operaio specializzato* e *Impiegato*.
- I punteggi invertono l'ordine di *Dirigenza* e *Professionista*.

Punteggi ed etichette dei valori

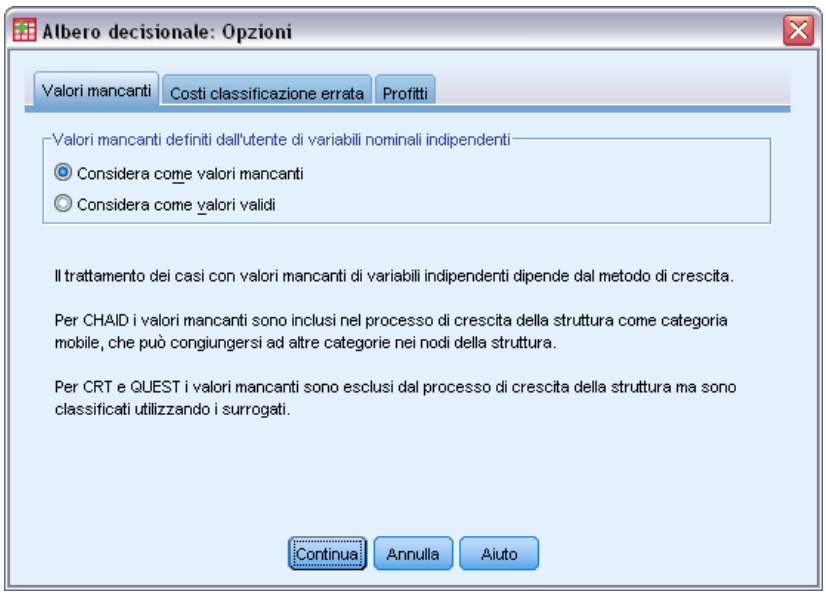
La finestra di dialogo richiede etichette dei valori definite per la variabile dipendente. Non è disponibile a meno che per almeno due valori della variabile dipendente categoriale siano state definite etichette dei valori.

Per specificare i punteggi

- Nella finestra di dialogo principale Albero decisionale, selezionare una variabile dipendente ordinale con due o più etichette dei valori definite.
- Come metodo di espansione scegliere CHAID o CHAID esaustivo.
- Fare clic su Opzioni.
- Fare clic sulla scheda Punteggi.

Valori mancanti

Figura 1-17
Finestra di dialogo Opzioni, scheda Valori mancanti



La scheda Valori mancanti controlla la gestione dei valori delle variabili nominali mancanti definiti dall'utente, indipendenti (predittore).

- La gestione dei valori delle variabili indipendenti definite dall'utente di scala e ordinali varia in base ai metodi di espansione.
- La gestione delle variabili dipendenti nominali è specificata nella finestra di dialogo Categorie. Per ulteriori informazioni, vedere l'argomento Selezione delle categorie a pag. 6.
- Per le variabili dipendenti di scala e ordinali i casi con valori di variabili dipendenti mancanti di sistema o mancanti definiti dall'utente vengono sempre esclusi.

Considera come valori mancanti. I valori mancanti definiti dall'utente sono considerati come mancanti di sistema. La gestione dei valori mancanti di sistema varia in base ai metodi di espansione.

Considera come valori validi. I valori mancanti definiti dall'utente di variabili indipendenti nominali sono considerati come valori ordinari nell'espansione dell'albero e nella classificazione.

Regole dipendenti dal metodo

Se alcuni, ma non tutti, i valori delle variabili indipendenti sono mancanti di sistema o definiti dall'utente:

- per CHAID e CHAID esaustivo, i valori mancanti definiti dall'utente e di sistema per le variabili indipendenti sono inclusi nell'analisi come una categoria singola combinata. Per le variabili indipendenti ordinali e di scala, gli algoritmi prima generano le categorie utilizzando i valori validi, quindi stabiliscono se unire la categoria mancante alla categoria (valida) più simile o se mantenerla separata.
- Per CRT e QUEST, i casi con i valori delle variabili indipendenti mancanti sono esclusi dal processo di espansione dell'albero ma sono classificati utilizzando i surrogati se i surrogati sono inclusi nel metodo. Se i valori mancanti definiti dall'utente nominali sono considerati come mancanti, vengono anch'essi gestiti nello stesso modo. Per ulteriori informazioni, vedere l'argomento Surrogati a pag. 16.

Per specificare il trattamento dei valori mancanti definiti dall'utente indipendenti nominali

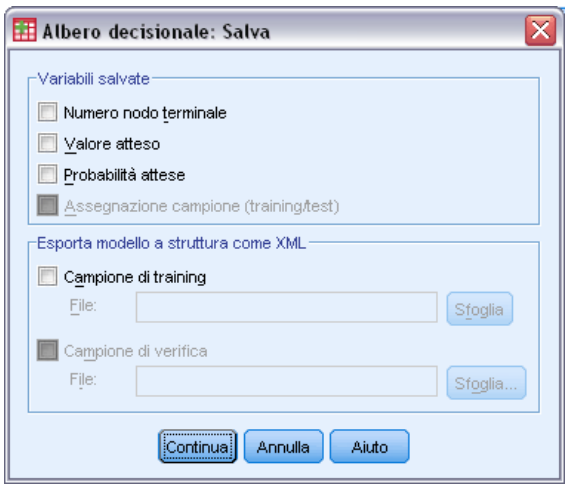
▶ Nella finestra di dialogo principale Albero decisionale selezionare almeno una variabile indipendente nominale.

▶ Fare clic su Opzioni.

▶ Fare clic sulla scheda Valori mancanti.

Salvataggio delle informazioni del modello

Figura 1-18
Salva



È possibile salvare le informazioni dal modello come variabili nel file dati di lavoro, oltreché salvare l'intero modello in formato XML (PMML) in un file esterno.

Variabili salvate

Numero dei nodi terminali. Il nodo terminale cui è assegnato ciascun caso. Il valore è il numero dei nodi dell'albero.

Valore atteso. La classe (gruppo) o valore per la variabile dipendente previsto dal modello.

Probabilità previste. La probabilità associata alla previsione del modello. Viene salvata una variabile per ogni categoria della variabile dipendente. Non disponibile per variabili dipendenti di scala.

Assegnazione di campioni (addestramento/verifica). Per la convalida con suddivisione, la variabile indica se un caso è stato utilizzato nel campione di verifica o di addestramento. Il valore è 1 per il campione di addestramento e 0 per il campione di verifica. Non disponibile a meno che sia stata selezionata la convalida con suddivisione. Per ulteriori informazioni, vedere l'argomento Convalida a pag. 8.

Esporta modello a struttura come XML

È possibile salvare il modello dell'intero albero in formato XML (PMML). È possibile utilizzare questo file di modello per applicare le informazioni del modello ad altri file di dati per il calcolo del punteggio.

Campione di addestramento. Scrive il modello nel file specificato. Per gli alberi convalidati con suddivisione, è il modello per il campione di addestramento.

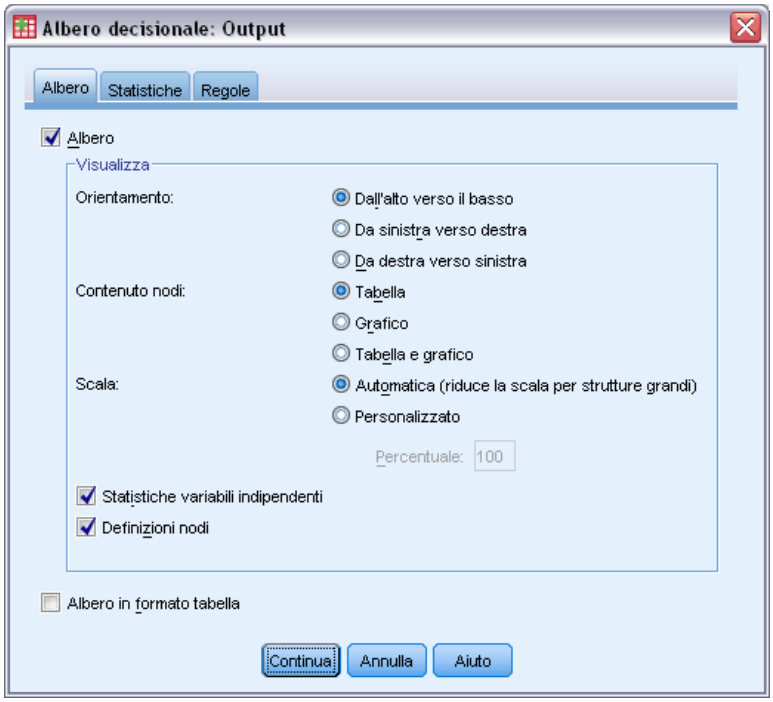
Campione di verifica. Scrive il modello per il campione di verifica nel file specificato. Non disponibile a meno che sia stata selezionata la convalida con suddivisione.

Output

Le opzioni di output disponibili dipendono dal metodo di espansione, dal livello di misurazione della variabile dipendente e da altre impostazioni.

Visualizzazione dell'albero

Figura 1-19
Finestra di dialogo Output, scheda Albero



È possibile controllare l'aspetto iniziale dell'albero o eliminarne completamente la visualizzazione.

Albero. Per impostazione predefinita, il diagramma ad albero è incluso nell'output visualizzato nel Viewer. Deselezionare questa opzione per escludere il diagramma ad albero dall'output.

Visualizzazione. Le opzioni controllano l'aspetto iniziale del diagramma nel Viewer. Tutti questi attributi possono inoltre essere modificati modificando l'albero generato.

- **Orientamento.** L'albero può essere visualizzato dall'alto in basso con il nodo radice in alto, da sinistra a destra o da destra a sinistra.
- **Contenuto dei nodi.** I nodi possono visualizzare tabelle, grafici o entrambi. Per le variabili dipendenti categoriali, le tabelle visualizzano conteggi di frequenza e percentuali; i grafici sono grafici a barre. Per le variabili dipendenti di scala, le tabelle visualizzano medie, deviazioni standard, numero di casi e valori attesi; i grafici sono istogrammi.
- **Scala.** Per impostazione predefinita, gli alberi di grandi dimensioni sono ridotti automaticamente per tentare di adattare l'albero alla pagina. È possibile specificare una percentuale di scala personalizzata fino al 200%.

- **Statistiche di variabili indipendenti.** Per CHAID e CHAID esaustivo, le statistiche includono il valore *F* (per le variabili dipendenti di scala) o il valore chi-quadrato (per le variabili dipendenti categoriali), nonché il valore di significatività e i gradi di libertà. Per CRT, il valore di miglioramento è indicato. Per QUEST *F*, il valore di significatività e i gradi di libertà sono indicati per le variabili indipendenti di scala e ordinali; per le variabili indipendenti nominali, sono indicati chi-quadrato, valore di significatività e gradi di libertà.
- **Definizioni dei nodi** Le definizioni dei nodi visualizzano il valore o i valori della variabile indipendente utilizzata per ciascuna divisione di nodo.

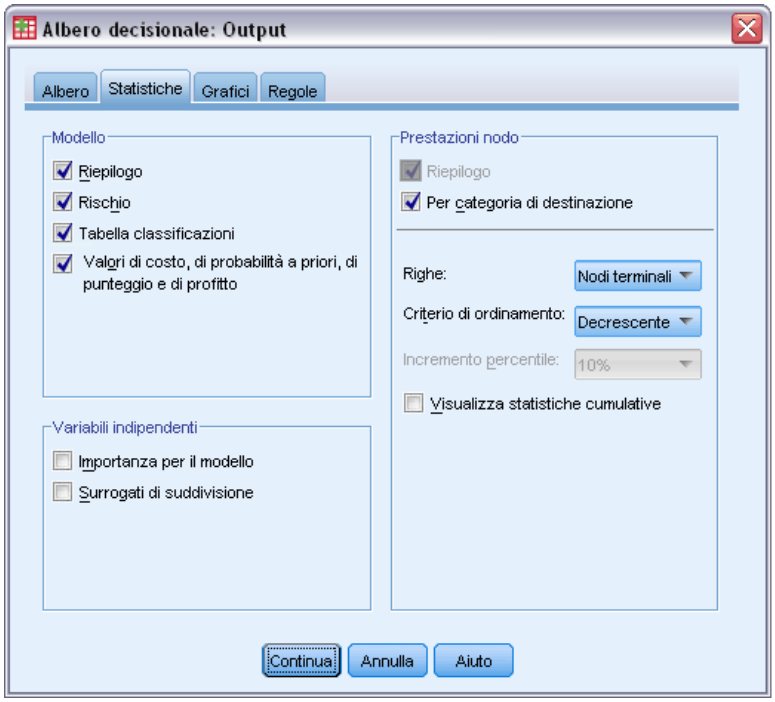
Albero in formato tabella. Informazioni riassuntive per ciascun nodo dell'albero, compresi numero del nodo genitore, statistiche della variabile dipendente, valore o valori della variabile indipendente per il nodo, media e deviazione standard per le variabili dipendenti di scala, oppure conteggi e percentuali per le variabili dipendenti categoriali.

Figura 1-20
Albero in formato tabella

	Basso		Alto		Totale				Primary Independent Variable				
Nodo	N	Percentuale	N	Percentuale	N	Percentuale	Categoria prevista	Nodo padre	Variabile	Correzione per confronti multipli[a]	Chi-quadrato	df	Valori divisione
0	1020	41,4%	1444	58,6%	2464	100,0%	Alto						
1	454	82,1%	99	17,9%	553	22,4%	Basso	0	Livello di reddito	,000	662,457	2	<= Basso
2	476	42,0%	658	58,0%	1134	46,0%	Alto	0	Livello di reddito	,000	662,457	2	(Basso, Medio]
3	90	11,6%	687	88,4%	777	31,5%	Alto	0	Livello di reddito	,000	662,457	2	> Medio
4	422	56,7%	322	43,3%	744	30,2%	Basso	2	Numero di carte di credito	,000	193,113	1	5 o più
5	54	13,8%	336	86,2%	390	15,8%	Alto	2	Numero di carte di credito	,000	193,113	1	Meno di 5
6	80	17,6%	375	82,4%	455	18,5%	Alto	3	Numero di carte di credito	,000	38,587	1	5 o più
7	10	3,1%	312	96,9%	322	13,1%	Alto	3	Numero di carte di credito	,000	38,587	1	Meno di 5
8	211	80,8%	50	19,2%	261	10,6%	Basso	4	Età	,000	95,299	1	<= 28,079205 81899067 6
9	211	43,7%	272	56,3%	483	19,6%	Alto	4	Età	,000	95,299	1	> 28,079205 81899067 6

Statistiche

Figura 1-21
Finestra di dialogo Output, scheda Statistiche



Le tabelle delle statistiche disponibili dipendono dal livello di misurazione della variabile dipendente, dal metodo di espansione e da altre impostazioni.

Modello

Tabella riassuntiva La tabella riassuntiva include il metodo utilizzato, le variabili incluse nel modello e le variabili specificate ma non incluse nel modello.

Figura 1-22
Tabella Riepilogo del modello

Specifiche	Metodo di crescita	CHAID
	Variabile dipendente	Merito di credito
	Variabili indipendenti	Età, Livello di reddito, Numero di carte di credito, Istruzione, Prestiti auto
	Convalida	NONE
	Massima profondità struttura	3
	Numero minimo di casi nel nodo padre	400
	Numero minimo di casi nel nodo figlio	200
Risultati	Variabili indipendenti incluse	Livello di reddito, Numero di carte di credito, Età
	Numero di nodi	10
	Numero di nodi terminali	6
	Profondità	3

Rischio. Stima del rischio e relativo errore standard. Una misura della precisione predittiva dell'albero.

- Per variabili dipendenti categoriali, la stima del rischio è la proporzione di casi erroneamente classificati dopo la correzione in base alle probabilità a priori e ai costi di errata classificazione.
- Per le variabili dipendenti di scala, la stima del rischio è la varianza all'interno del nodo.

Tabella classificazioni. Per le variabili dipendenti categoriali (nominali, ordinali) la tabella mostra il numero dei casi classificati correttamente e non per ciascuna categoria della variabile dipendente. Non disponibile per variabili dipendenti di scala.

Figura 1-23
Rischio e tabelle di classificazione

Rischio

Stima	Errore standard
,275	,008

Metodo di crescita: CHAID
Variabile dipendente: Merito di credito

Classificazione

Osservato	Previsione		
	Basso	Alto	Percentuale corretta
Basso	665	355	65,2%
Alto	149	1295	89,7%
Percentuale globale	33,0%	67,0%	79,5%

Metodo di crescita: CHAID
Variabile dipendente: Merito di credito

Valori di costo, probabilità a priori e profitto Per le variabili dipendenti categoriali la tabella mostra i valori di costo, probabilità a priori, punteggio e profitto utilizzati nell'analisi. Non disponibile per variabili dipendenti di scala.

Variabili indipendenti

Importanza per il modello Per il metodo di espansione CRT, classifica ogni variabile (predittore) indipendente in base alla sua importanza per il modello. Non disponibile per i metodi QUEST o CHAID.

Surrogati di suddivisione. Per i metodi di espansione CRT e QUEST, se il modello include surrogati, elenca i surrogati per ciascuna divisione nell'albero. Non disponibile per i metodi CHAID. Per ulteriori informazioni, vedere l'argomento Surrogati a pag. 16.

Prestazioni nodo

Tabella riassuntiva Per variabili dipendenti di scala, la tabella include il numero di nodi, il numero di casi e il valore della media della variabile dipendente. Per variabili dipendenti categoriali con profitti definiti, la tabella include i valori di numero di nodi, numero di casi, profitto medio e ROI

(return on investment). Non disponibile per variabili dipendenti categoriali senza profitti definiti. Per ulteriori informazioni, vedere l'argomento Profitti a pag. 18.

Figura 1-24
Tabelle riassuntive di guadagno per nodi e percentili

Riepilogo guadagni per nodi

Nodo	N	Percentuale	Profitto	Rendimento capitale investito
7	322	13,1%	77,826	377,4%
5	390	15,8%	70,308	308,8%
6	455	18,5%	67,692	287,9%
9	483	19,6%	49,420	172,0%
8	261	10,6%	23,410	64,7%
1	553	22,4%	22,532	61,9%

Riepilogo guadagni per percentili

Percentile	Nodi	N	Profitto	Rendimento capitale investito
10	7	246	77,826	377,4%
20	7 ; 5	493	75,218	352,0%
30	5 ; 6	739	73,488	336,2%
40	6	986	72,036	323,4%
50	6 ; 9	1232	70,205	307,9%
60	9	1478	66,745	280,6%
70	9 ; 8	1725	63,134	254,4%
80	8 ; 1	1971	58,149	2216,6%
90	1	2218	54,183	197,9%
100	1	2464	51,023	180,4%

Per categoria obbiettivo. Per variabili dipendenti categoriali con categorie obbiettivo definite, la tabella include il guadagno in percentuale, la percentuale di risposta e la percentuale dell'indice (lift) per nodo o gruppo di percentili. Per ciascuna categoria obbiettivo verrà creata una tabella distinta. Non disponibile per variabili dipendenti di scala o categoriali senza categorie obbiettivo definite. Per ulteriori informazioni, vedere l'argomento Selezione delle categorie a pag. 6.

Figura 1-25
Guadagni di categorie obbiettivo per nodi e percentili

Target Category: Basso

Guadagni per nodi

Nodo	Nodo		Guadagno		Risposta	Indice
	N	Percentuale	N	Percentuale		
1	553	22,4%	454	44,5%	82,1%	198,3%
8	261	10,6%	211	20,7%	80,8%	195,3%
9	483	19,6%	211	20,7%	43,7%	105,5%
6	455	18,5%	80	7,8%	17,6%	42,5%
5	390	15,8%	54	5,3%	13,8%	33,4%
7	322	13,1%	10	1,0%	3,1%	7,5%

Guadagni per percentili

Percentile	Nodi	N	Guadagno		Risposta	Indice
			N	Percentuale		
10	17 ; 9	246	202	19,8%	82,1%	198,3%
20	9 ; 19	493	405	39,7%	82,1%	198,3%
30	19 ; 16	739	604	59,3%	81,8%	197,6%
40	16 ; 14	986	740	72,6%	75,1%	181,3%
50	14 ; 15 ; 18	1232	848	83,1%	68,8%	166,2%
60	18 ; 5 ; 13	1478	908	89,0%	61,4%	148,4%
70	13	1725	951	93,3%	55,1%	133,2%
80	13 ; 11 ; 12	1971	986	96,7%	50,0%	120,9%
90	12 ; 10	2218	1012	99,3%	45,6%	110,3%
100	10	2464	1020	100,0%	41,4%	100,0%

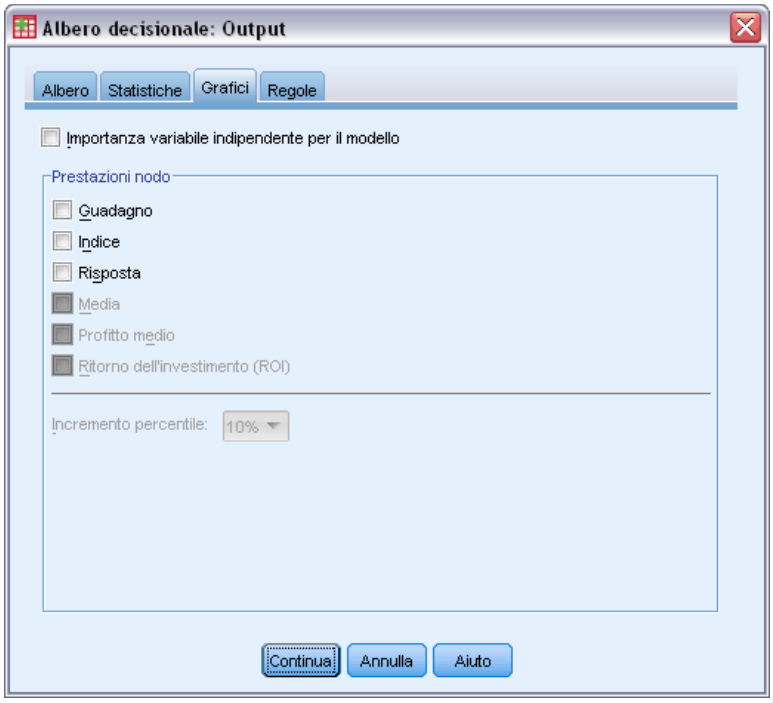
Righe Le tabelle delle prestazioni dei nodi possono visualizzare i risultati per nodi terminali, percentili o entrambi. Se si selezionano entrambi, per ciascuna categoria obbiettivo verranno create due tabelle. Le tabelle dei percentili visualizzano valori cumulati per ciascun percentile, on base all'ordinamento.

Incremento percentile. Per le tabelle di percentile, è possibile selezionare l'incremento di percentile: 1, 2, 5, 10, 20 o 25.

Visualizza statistiche cumulate. Per le tabelle dei nodi terminali, visualizza colonne aggiuntive in ciascuna tabella con risultati cumulati.

Grafici

Figura 1-26
Finestra di dialogo Output, scheda Grafici



I grafici disponibili dipendono dal livello di misurazione della variabile dipendente, dal metodo di espansione e da altre impostazioni.

Importanza variabile indipendente per il modello. Il grafico a barre dell'importanza del modello per variabile indipendente (predittore). Disponibile solo con il metodo di espansione CRT .

Prestazioni nodo

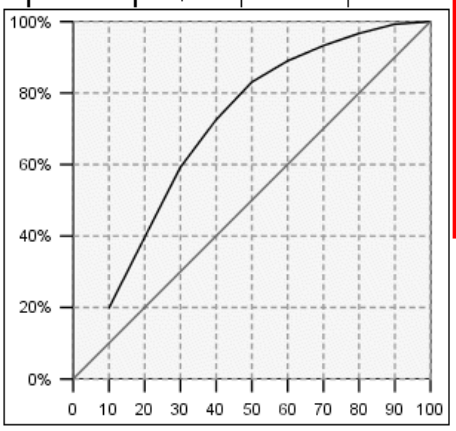
Guadagno. Il guadagno è la percentuale dei casi totali nella categoria obbiettivo in ciascun nodo, calcolato come segue: (obbiettivo nodo*n*/obbiettivo totale *n*) x 100. Il grafico dei guadagni è una grafico lineare dei guadagni percentili cumulati, calcolato come segue: (obbiettivo percentile cumulato *n*/obbiettivo totale *n*) x 100. Un grafico lineare separato viene prodotto per ciascuna categoria obbiettivo. Disponibile solo per variabili dipendenti categoriali con categorie obbiettivo definite. Per ulteriori informazioni, vedere l'argomento Selezione delle categorie a pag. 6.

Il grafico dei guadagni include gli stessi valori che sarebbero visualizzati nella colonna *Percentuale guadagno* nella +++tabella guadagni per percentili, che riporta anche i valori cumulati.

Figura 1-27
Guadagni per tabella dei percentili e grafico dei guadagni

Guadagni per percentili

Percentile	Nodi	N	Guadagno N	Guadagno Percentuale	Risposta	Indice
10	17 ; 9	246	202	19,8%	82,1%	198,3%
20	9 ; 19	493	405	39,7%	82,1%	198,3%
30	19 ; 16	739	604	59,3%	81,8%	197,6%
40	16 ; 14	986	740	72,6%	75,1%	181,3%
				83,1%	68,8%	166,2%
				89,0%	61,4%	148,4%
				93,3%	55,1%	133,2%
				96,7%	50,0%	120,9%
				99,3%	45,6%	110,3%
				100,0%	41,4%	100,0%



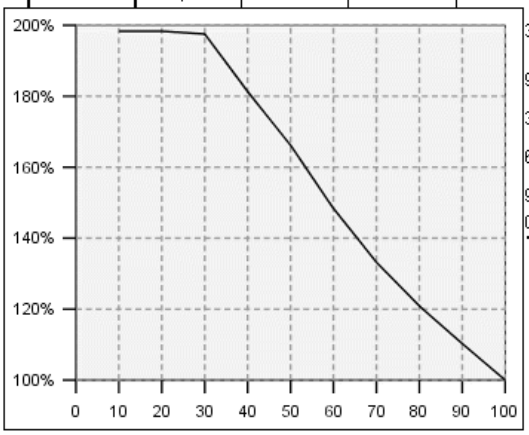
Indice. L'indice è il rapporto fra la percentuale di risposta del nodo per la categoria di destinazione e la percentuale di risposta globale per la categoria di destinazione dell'intero campione. Il grafico degli indici è un grafico lineare dei valori dell'indice dei percentili cumulati. Disponibile solo per variabili dipendenti categoriali. L'indice percentile cumulato è calcolato come segue: (percentuale risposta percentile cumulata / percentuale risposta totale) x 100. Un grafico separato viene prodotto per ciascuna categoria obbiettivo; è necessario che le categorie obbiettivo siano definite.

Il grafico degli indici include gli stessi valori che sarebbero visualizzati nella colonna *Indice* nella tabella guadagni per percentili.

Figura 1-28
Guadagni per tabella dei percentili e grafico degli indici

Guadagni per percentili

Percentile	Nodi	N	Guadagno N	Guadagno Percentuale	Risposta	Indice
10	17 ; 9	246	202	19,8%	82,1%	198,3%
20	9 ; 19	493	405	39,7%	82,1%	198,3%
30	19 ; 16	739	604	59,3%	81,8%	197,6%
40	16 ; 14	986	740	72,6%	75,1%	181,3%
				3,1%	68,8%	166,2%
				9,0%	61,4%	148,4%
				3,3%	55,1%	133,2%
				6,7%	50,0%	120,9%
				9,3%	45,6%	110,3%
				0,0%	41,4%	100,0%



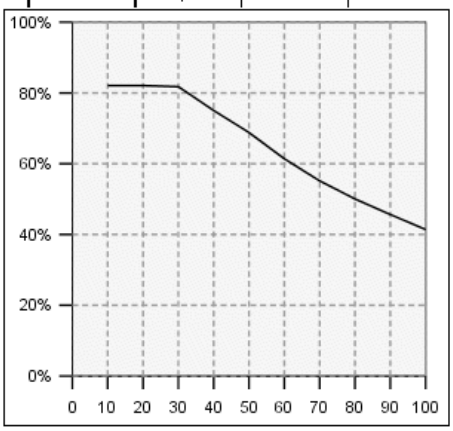
Risposta. La percentuale di casi nel nodo nella categoria di destinazione specificata. Il grafico delle risposte è un grafico lineare di risposta percentile cumulata, calcolata come segue: (obbiettivo percentile cumulato *n*/totale percentile cumulato *n*) x 100. Disponibile solo per le variabili dipendenti categoriali con categorie obbiettivo definite.

Il grafico delle risposte include gli stessi valori che sarebbero visualizzati nella colonna *Risposta* nella tabella guadagni per percentili.

Figura 1-29
Guadagni per tabella dei percentili e grafico delle risposte

Guadagni per percentili

Percentile	Nodi	N	Guadagno N	Guadagno Percentuale	Risposta	Indice
10	17 ; 9	246	202	19,8%	82,1%	198,3%
20	9 ; 19	493	405	39,7%	82,1%	198,3%
30	19 ; 16	739	604	59,3%	81,8%	197,6%
40	16 ; 14	986	740	72,6%	75,1%	181,3%
				83,1%	68,8%	166,2%
				89,0%	61,4%	148,4%
				93,3%	55,1%	133,2%
				96,7%	50,0%	120,9%
				99,3%	45,6%	110,3%
				100,0%	41,4%	100,0%



Media. Grafico lineare dei valori delle medie percentili cumulate per la variabile dipendente: Disponibile solo per variabili dipendenti di scala.

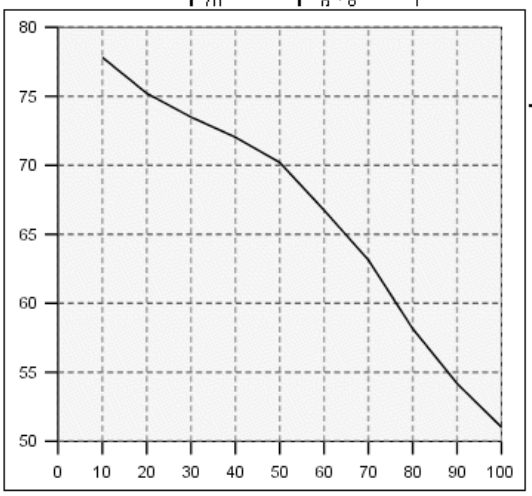
Profitto medio. Grafico lineare del profitto medio cumulato. Disponibile solo per variabili dipendenti categoriali con profitti definiti. Per ulteriori informazioni, vedere l'argomento Profitti a pag. 18.

Il grafico dei profitti medi include gli stessi valori che sarebbero visualizzati nella colonna *Profitto* nella tabella riepilogo guadagni per percentili.

Figura 1-30
Riepilogo guadagni per tabella dei percentili e grafico dei profitti medi

Riepilogo guadagni per percentili

Percentile	Nodi	N	Profitto	Rendimento capitale investito
10	7	246	77,826	377,4%
20	7 ; 5	493	75,218	352,0%
30	5 ; 6	739	73,488	336,2%
40	6	986	72,036	323,4%
50	6 ; 9	1232	70,205	307,9%
60	9	1478	66,745	280,6%
70	9 ; 8		63,134	254,4%
			58,149	2216,6%
			54,183	197,9%
			51,023	180,4%



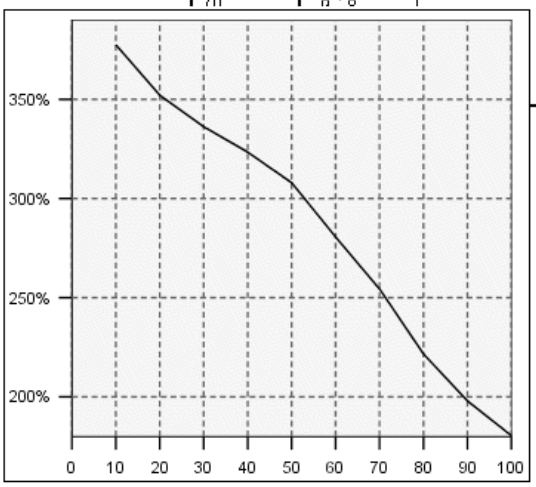
Return on investment (ROI). Grafico lineare del ROI (return on investment) cumulato. Il ROI è calcolato come il rapporto tra profitti e spese. Disponibile solo per variabili dipendenti categoriali con profitti definiti.

Il grafico del ROI include gli stessi valori che sarebbero visualizzati nella colonna *ROI* nella tabella riepilogo guadagni per percentili.

Figura 1-31
Riepilogo guadagni per tabella dei percentili e grafico del ROI

Riepilogo guadagni per percentili

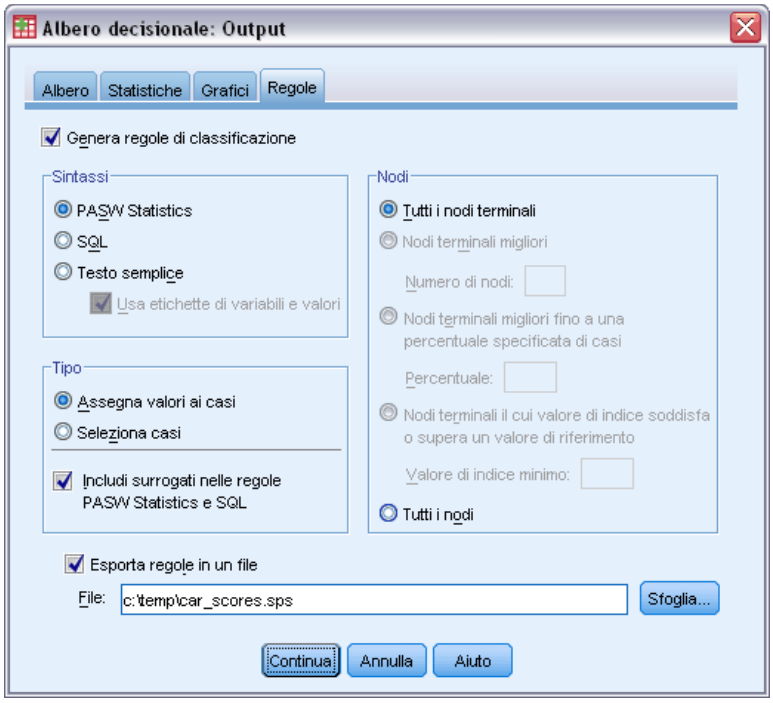
Percentile	Nodi	N	Profitto	Rendimento capitale investito
10	7	246	77,826	377,4%
20	7 ; 5	493	75,218	352,0%
30	5 ; 6	739	73,488	336,2%
40	6	986	72,036	323,4%
50	6 ; 9	1232	70,205	307,9%
60	9	1478	66,745	280,6%
70	9 ; 8		63,134	254,4%
			58,149	2216,6%
			54,183	197,9%
			51,023	180,4%



Incremento percentile. Per tutti i grafici dei percentili, questa impostazione controlla gli incrementi dei percentili visualizzati nel grafico: 1, 2, 5, 10, 20 o 25.

Regole di selezione e di punteggio

Figura 1-32
Finestra di dialogo Output, scheda Regole



La scheda Regola consente di generare regole di previsione/classificazione o selezione sotto forma di sintassi di comando, SQL o formato testo standard (Inglese). È possibile visualizzare tali regole nel Viewer e/o salvarle in un file esterno.

Sintassi. Controlla la forma delle regole di selezione nell'output visualizzato nel Viewer e/o nelle regole di selezione salvate in un file esterno.

- **IBM® SPSS® Statistics** Linguaggio della sintassi dei comandi. Le regole sono espresso come un insieme di comandi che definiscono una condizione di filtro utilizzabile per selezionare sottoinsiemi di casi o come dichiarazioni `COMPUTE` utilizzabili per assegnare punteggi ai casi.
- **SQL.** Le regole SQL standard sono generate per selezionare o estrarre record da un database o assegnare valori a tali record. Le regole SQL generate non includono nomi di tabella o altre informazioni sulle origini dati.
- **Testo semplice.** Pseudo-codice in inglese standard Le regole sono espresse come insieme di dichiarazioni logiche "if...then" che descrivono le classificazioni del modello o le previsioni per ciascun nodo. Le regole con questo formato possono utilizzare etichette dei valori o di variabile definite oppure nomi delle variabili o valori di dati.

Tipo. Per le regole SPSS Statistics e SQL, controlla il tipo di regole generate: regole di selezione o di punteggio.

- **Assegna valori a casi.** Le regole possono essere utilizzate per assegnare le previsioni del modello a casi che rispondono ai criteri di appartenenza del nodo. Una regola separata viene generata per ciascun nodo che risponde ai criteri di appartenenza del nodo.
- **Seleziona casi.** Le regole possono essere utilizzate per selezionare casi che rispondono ai criteri di appartenenza del nodo. Per le regole SPSS Statistics e SQL, una regola singola viene generata per selezionare tutti i casi che rispondono ai criteri di selezione.

Includi surrogati nelle regole SPSS Statistics e SQL. Per CRT e QUEST è possibile includere nelle regole predittori di surrogati dal modello. Le regole che includono surrogati sono alquanto complesse. In generale, se si desidera semplicemente ricavare informazioni concettuali sull'albero, escludere i surrogati. Se per alcuni casi i dati (predittore) della variabile indipendente sono incompleti e si desiderano regole che simulino l'albero, includere i surrogati. Per ulteriori informazioni, vedere l'argomento Surrogati a pag. 16.

Nodi. Controlla l'ambito delle regole generate. Per ogni nodo incluso nell'ambito viene creata una regola distinta.

- **Tutti i nodi terminali.** Genera regole per ogni nodo terminale.
- **Nodi terminali migliori** Genera regole per i primi *n* nodi terminali in base ai valori dell'indice. Se il numero supera quello dei nodi terminali dell'albero, le regole vengono generate per tutti i nodi terminali (vedere la nota seguente).
- **Nodi terminali migliori fino a una percentuale di casi specificata.** Genera regole per i nodi terminali per la percentuale dei primi *n* casi in base ai valori dell'indice. (vedere la nota seguente).
- **Nodi terminali il cui valore di indice è uguale o supera un valore di riferimento.** Genera regole per tutti i nodi terminali con valore di indice maggiore o uguale al valore specificato. Un valore di indice maggiore di 100 significa che la percentuale di casi nella categoria obbiettivo del nodo è maggiore rispetto alla percentuale nel nodo radice. (vedere la nota seguente).
- **Tutti i nodi.** Genera regole per tutti i nodi.

Nota 1: la selezione dei nodi in base ai valori dell'indice è disponibile solo per variabili dipendenti categoriali con categorie obbiettivo definite. Se sono state specificate categorie obbiettivo multiple, viene generato un insieme separato di regole per ogni categoria obbiettivo.

Nota 2: per le regole SPSS Statistics e SQL per la selezione di casi (non per l'assegnazione di valori), selezionando Tutti i nodi e Tutti i nodi terminali verrà generata una regola che selezionerà tutti i casi utilizzati nell'analisi.

Esporta regole in un file. Salva le regole in un file di testo esterno.

È inoltre possibile generare e salvare regole per la selezione o l'assegnazione di punteggio in modo interattivo, in base a nodi selezionati nel modello di albero finale. Per ulteriori informazioni, vedere l'argomento Regole di selezione e di punteggio dei casi in il capitolo 2 a pag. 46.

Nota: se si applicano le regole sotto forma di sintassi di comando a un altro file dati, questo dovrà contenere variabili con gli stessi nomi delle variabili indipendenti incluse nel modello finale, misurate nella stessa metrica e con gli stessi valori mancanti definibili dall'utente (se presenti).

Capitolo
2

Editor albero

Nell'Editor degli alberi è possibile:

- Visualizzare o nascondere rami selezionati.
- Controllare la visualizzazione del contenuto del nodo, delle statistiche visualizzate per le divisioni dei nodi e di altre informazioni.
- Modificare nodo, sfondo, bordo, grafico e colore dei caratteri.
- Modificare stile e dimensione dei caratteri.
- Modificare l'allineamento dell'albero.
- Selezionare sottoinsiemi di casi per un'ulteriore analisi in base a nodi selezionati.
- Creare e salvare regole per la selezione o l'assegnazione di punteggio ai casi in base a nodi selezionati.

Per modificare un modello ad albero:

▶ Fare doppio clic sul modello nella finestra Viewer.

o

▶ Dal menu Modifica o dal menu di scelta rapida scegliere:
Modifica contenuto > In una finestra separata

Nascondere e visualizzare i nodi

Per nascondere (comprimere) tutti i nodi figlio di un ramo di livello inferiore a un nodo genitore:

▶ fare clic sul segno meno (–) nella casellina sotto l'angolo inferiore destro del nodo genitore.

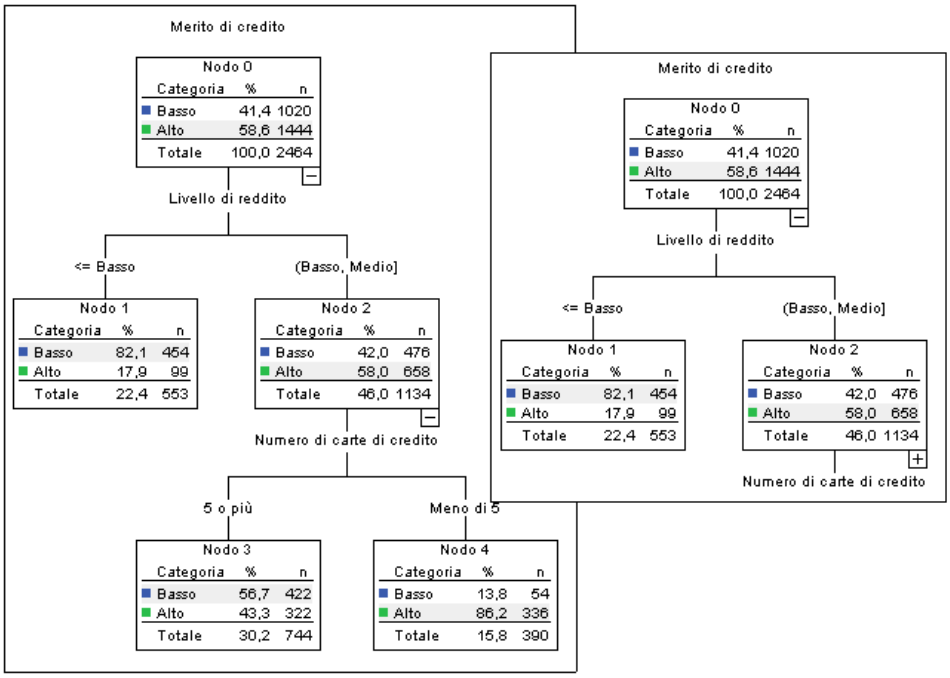
Tutti i nodi sotto il nodo genitore sul ramo verranno nascosti.

Per visualizzare (espandere) tutti i nodi figlio di un ramo di livello inferiore a un nodo genitore:

▶ fare clic sul segno più (-) nella casellina sotto l'angolo inferiore destro del nodo genitore.

Nota: nascondere i nodi figlio di un ramo non è equivalente a tagliare l'albero. Se si desidera un albero tagliato, è necessario richiedere il taglio prima di creare l'albero e i rami tagliati non vengono inclusi nell'albero finale. Per ulteriori informazioni, vedere l'argomento Taglio degli alberi in il capitolo 1 a pag. 15.

Figura 2-1
Albero espanso e compresso



Selezione di più nodi

È possibile selezionare casi, generare regole per la selezione o l'assegnazione di punteggio ed eseguire altre operazioni in base al nodo o ai nodi selezionati. Per selezionare più nodi:

- Fare clic sul nodo da selezionare.
- Fare clic sugli altri nodi da selezionare tenendo premuto il tasto Ctrl.

È possibile selezionare più nodi fratello e/o genitore in un ramo e nodi figlio in un altro. Non è possibile tuttavia applicare la selezione multipla a un nodo genitore e a un nodo figlio/discendente dello stesso ramo.

Utilizzo di alberi di grandi dimensioni

I modelli ad albero possono a volte includere un numero tale di nodi e di rami da rendere difficile o impossibile la visualizzazione dell"intero albero. Esistono varie funzioni utili quando si utilizzano alberi di grandi dimensioni:

- **Mappa dell'albero.** È possibile utilizzare la mappa dell'albero, una versione ridotta e semplificata dell'albero, per spostarsi all'interno dell'albero e selezionare i nodi. Per ulteriori informazioni, vedere l'argomento Mappa albero a pag. 41.

- **Scaling.** È possibile applicare lo zoom avanti e indietro modificato la percentuale di scale della visualizzazione dell'albero. Per ulteriori informazioni, vedere l'argomento Scaling della visualizzazione dell'albero a pag. 42.
- **Visualizzazione di rami e nodi.** È possibile rendere un albero più compatto visualizzando solo le tabelle o i grafici nei nodi e/o eliminando la visualizzazione delle etichette dei nodi o delle informazioni sulle variabili indipendenti. Per ulteriori informazioni, vedere l'argomento Controllo delle informazioni visualizzate nell'albero a pag. 43.

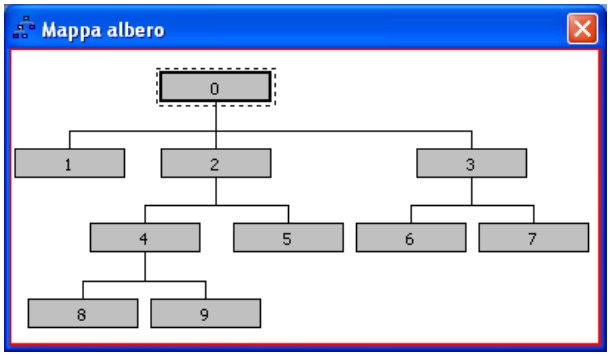
Mappa albero

La mappa dell'albero fornisce una versione ridotta e semplificata dell'albero, utilizzabile per spostarsi all'interno dell'albero e per selezionare i nodi.

Per utilizzare la finestra della mappa dell'albero:

▶ Dai menu dell'Editor degli alberi, scegliere:
Visualizza > Mappa albero

Figura 2-2
Finestra della mappa dell'albero



- Il nodo attualmente selezionato è evidenziato sia nell'Editor del modello ad albero sia nella finestra della mappa.
- L'area dell'albero attualmente visualizzata nell'area di visualizzazione dell'Editor è indicata da un rettangolo rosso nella mappa dell'albero. Fare clic con il pulsante destro del mouse e trascinare il rettangolo per modificare la sezione dell'albero visualizzata nell'area di visualizzazione.
- Se si seleziona un nodo nella mappa dell'albero che attualmente non è compreso nell'area di visualizzazione dell'Editor, la visualizzazione si modifica in modo da includere il nodo selezionato.
- La selezione di nodi multipli funziona in modo analogo nella mappa dell'albero e nell'Editor: fare clic tenendo premuto il tasto Ctrl per selezionare più nodi. Non è possibile applicare la selezione multipla a un nodo genitore e a un nodo figlio/discendente dello stesso ramo.

Scaling della visualizzazione dell'albero

Per impostazione predefinita, gli alberi vengono scalati automaticamente per adattarsi alla pagina del Viewer, il che può determinarne inizialmente una certa difficoltà di lettura. È possibile selezionare un'impostazione predefinita di scala oppure creare un proprio valore personalizzato compreso tra 5% e 200%.

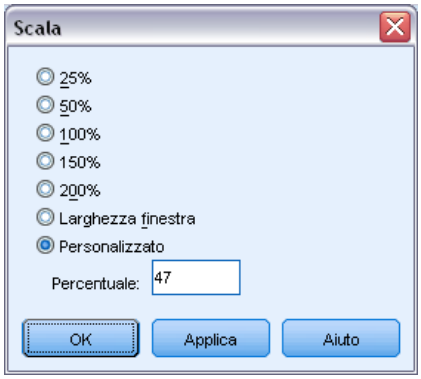
Per modificare la scala dell'albero:

- Selezionare una percentuale di scala dall'elenco a discesa sulla barra degli strumenti o inserire un valore di percentuale personalizzato.

o

- Dai menu dell'Editor degli alberi, scegliere:
  Visualizza > Scala...

Figura 2-3
Finestra di dialogo scala



È inoltre possibile specificare un valore di scala prima di creare il modello ad albero. Per ulteriori informazioni, vedere l'argomento Output in il capitolo 1 a pag. 25.

Finestra Riepilogo nodi

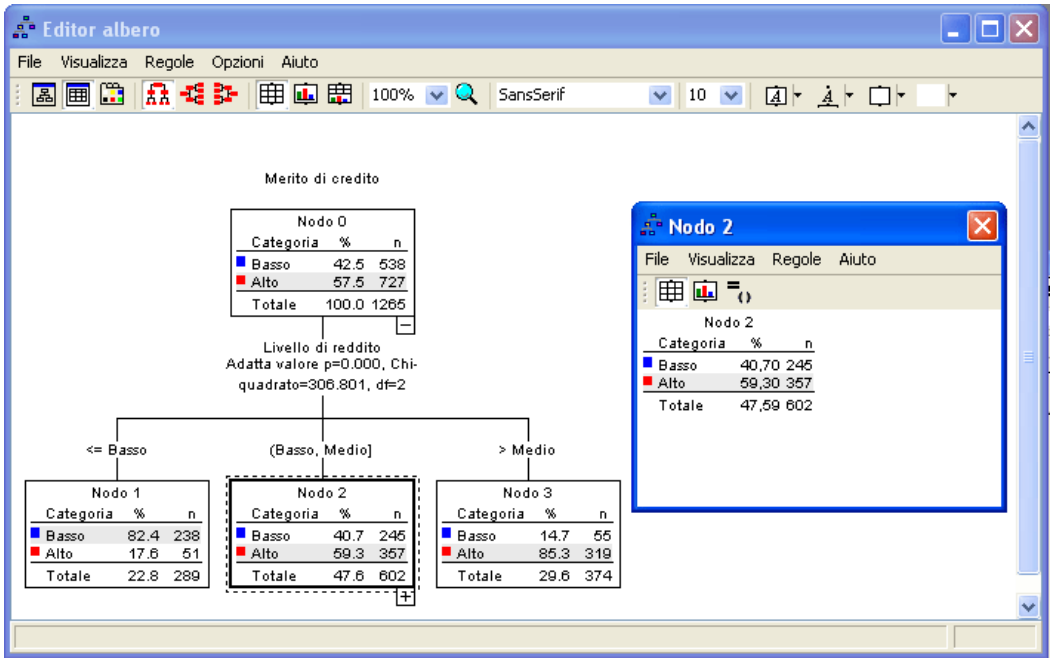
La finestra di riepilogo nodi offre una visualizzazione più ampia dei nodi selezionati. È inoltre possibile utilizzarla per visualizzare, applicare o salvare regole per la selezione o l'assegnazione di punteggio basate sui nodi selezionati.

- Utilizzare il menu Visualizza nella finestra di riepilogo nodi per spostarsi tra le visualizzazioni di una tabella riassuntiva, del grafico o delle regole.
- Utilizzare il menu Regole nella finestra di riepilogo nodi per selezionare il tipo di regole da visualizzare. Per ulteriori informazioni, vedere l'argomento Regole di selezione e di punteggio dei casi a pag. 46.
- Tutte le visualizzazioni nella finestra di riepilogo nodi offrono un riepilogo combinato per tutti i nodi selezionati.

Per utilizzare la finestra di riepilogo nodi:

- Selezionare i nodi nell'Editor degli alberi. Fare clic tenendo premuto il tasto Ctrl per selezionare più nodi.
- Dai menu, scegliere:
  Visualizza > Riepilogo

Figura 2-4
Finestra Riepilogo

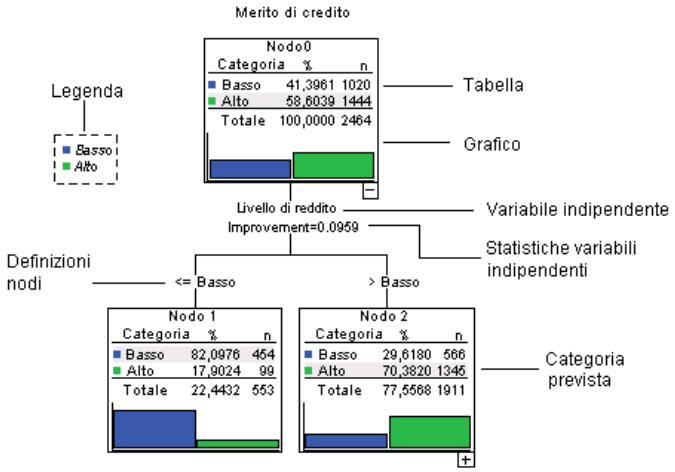


Controllo delle informazioni visualizzate nell'albero

Il menu Opzioni dell'Editor degli alberi consente di controllare la visualizzazione del contenuto dei nodi, i nomi e le statistiche delle variabili indipendenti (predittore), le definizioni dei nodi e altre impostazioni. Molte di queste impostazioni possono essere controllate anche dalla barra degli strumenti.

Impostazione	**Selezione del menu Opzioni**
Evidenzia categoria prevista evidenziata (variabile dipendente.categoriale)	Evidenzia previsioni
Tabelle e/o grafici nel nodo.	Contenuto nodo
Valori di testo del livello di significatività e valori *p*	Statistiche variabili indipendenti
Nomi di variabili (predittore) indipendenti	Variabili indipendenti
Nodi per il valore o i valori (predittore) indipendenti	Definizione nodi
Allineamento (dall'alto verso il basso, da sinistra a destra, da destra a sinistra)	Orientamento
Legenda dei grafici	Legenda

Figura 2-5
Elementi dell'albero



Modifica dei colori dell'albero e dei caratteri del testo

È possibile modificare i seguenti colori nell'albero:

- Bordo dei nodi, sfondo e colore del testo
- Colore del ramo e colore del testo del ramo
- Colore di sfondo dell'albero
- Colore di evidenziazione della categoria prevista (variabili dipendenti.categoriali)
- Colori dei grafici dei nodi

È inoltre possibile modificare tipo di carattere, stile e dimensione per tutto il testo nell'albero.

Nota: non è possibile modificare il colore o gli attributi del tipo di carattere per singoli nodi o rami. Le modifiche ai colori si applicano a tutti gli elementi dello stesso tipo, mentre le modifiche al tipo di caratteri (diverse dal colore) si applicano a tutti gli elementi dei grafici.

Per modificare i colori dell'albero e gli attributi dei caratteri del testo:

▶ Utilizzare la barra degli strumenti per modificare gli attributi del carattere per l'intero albero o i colori per i diversi elementi dell'albero. Le descrizioni degli strumenti consentono di visualizzare una descrizione di ciascun controllo sulla barra degli strumenti posizionandovi sopra il puntatore del mouse.

o

▶ Fare doppio clic in qualsiasi punto dell'Editor degli alberi per aprire la finestra Proprietà, oppure dai menu scegliere:
Visualizza > Proprietà

▶ Per bordo, ramo, sfondo del nodo, categoria prevista, e sfondo dell'albero, fare clic sulla scheda Colore.

- Per i colori e gli attributi del carattere, fare clic sulla scheda Testo.
- Per i colori dei grafici dei nodi, fare clic sulla scheda Grafici nodo.

Figura 2-6
Finestra Proprietà, scheda Colore

Figura 2-7
Finestra Proprietà, scheda Testo

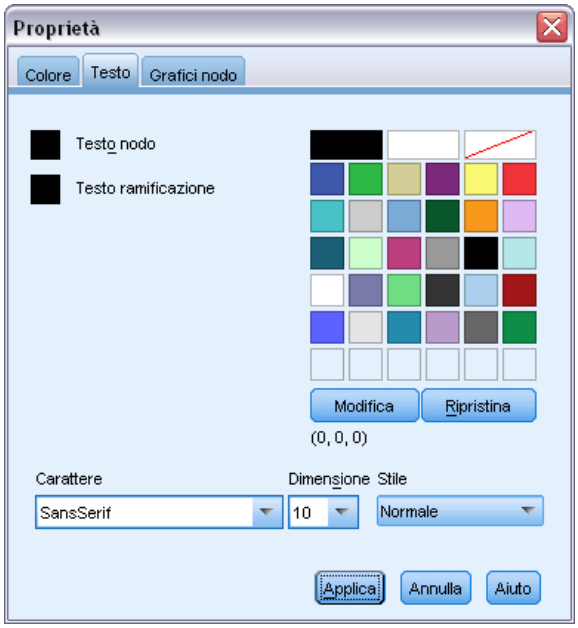
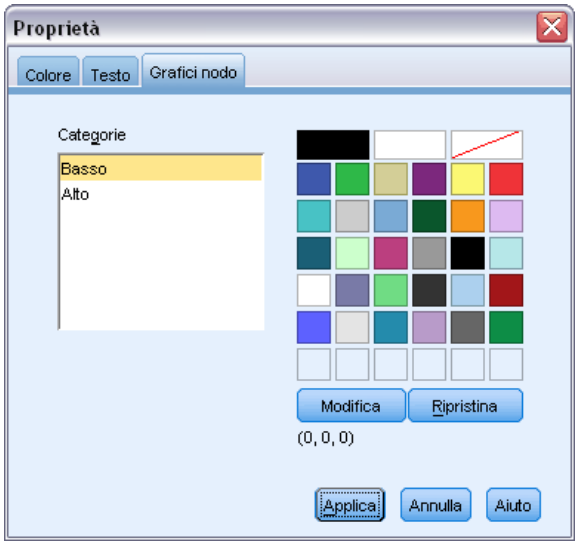


Figura 2-8
Finestra Proprietà, scheda Grafici nodo



Regole di selezione e di punteggio dei casi

È possibile utilizzare l'Editor degli alberi per:

- Selezionare sottoinsiemi di casi in base al nodo o ai nodi selezionati. Per ulteriori informazioni, vedere l'argomento Applicazione di filtri ai casi a pag. 46.
- Generare regole di selezione dei casi o di assegnazione di punteggio nel formato della sintassi dei comandi IBM® SPSS® Statistics o SQL. Per ulteriori informazioni, vedere l'argomento Salvataggio di regole di selezione e di punteggio a pag. 47.

È inoltre possibile salvare automaticamente le regole in base a vari criteri quando si esegue la procedura Albero decisionale per creare il modello dell'albero. Per ulteriori informazioni, vedere l'argomento Regole di selezione e di punteggio in il capitolo 1 a pag. 37.

Applicazione di filtri ai casi

Per maggiori informazioni sui casi in un particolare nodo o gruppo di nodi, selezionare un sottoinsieme di casi per un'ulteriore analisi basata sui nodi selezionati.

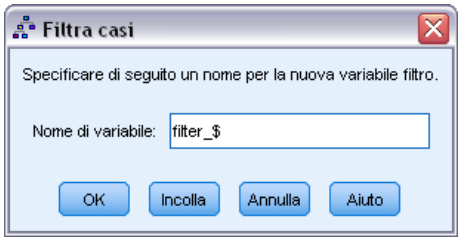
▶ Selezionare i nodi nell'Editor degli alberi. Fare clic tenendo premuto il tasto Ctrl per selezionare più nodi.

▶ Dai menu, scegliere:
Regole > Filtra casi...

▶ Inserire un nome per la variabile di filtro. I casi dei nodi selezionati riceveranno un valore pari a 1 per la variabile. Tutti gli altri casi riceveranno un valore pari a 0 e saranno esclusi dall'analisi successiva fino a quando non venga modificato lo stato del filtro.

▶ Fare clic su OK.

Figura 2-9
Finestra di dialogo Filtra casi



Salvataggio di regole di selezione e di punteggio

È possibile salvare le regole di selezione o di punteggio in un file esterno e quindi applicarle a una diversa origine dati. Le regole si basano sui nodi selezionati nell'Editor degli alberi.

Sintassi. Controlla la forma delle regole di selezione nell'output visualizzato nel Viewer e/o nelle regole di selezione salvate in un file esterno.

- **IBM® SPSS® Statistics** Linguaggio della sintassi dei comandi. Le regole sono espresso come un insieme di comandi che definiscono una condizione di filtro utilizzabile per selezionare sottoinsiemi di casi o come dichiarazioni `COMPUTE` utilizzabili per assegnare punteggi ai casi.
- **SQL.** Le regole SQL standard sono generate per selezionare o estrarre record da un database o assegnare valori a tali record. Le regole SQL generate non includono nomi di tabella o altre informazioni sulle origini dati.

Tipo. È possibile creare regole di selezione o di punteggio.

- **Seleziona casi.** Le regole possono essere utilizzate per selezionare casi che rispondono ai criteri di appartenenza del nodo. Per le regole SPSS Statistics e SQL, una regola singola viene generata per selezionare tutti i casi che rispondono ai criteri di selezione.
- **Assegna valori a casi.** Le regole possono essere utilizzate per assegnare le previsioni del modello a casi che rispondono ai criteri di appartenenza del nodo. Una regola separata viene generata per ciascun nodo che risponde ai criteri di appartenenza del nodo.

Includi surrogati. Per CRT e QUEST è possibile includere nelle regole predittori di surrogati dal modello. Le regole che includono surrogati sono alquanto complesse. In generale, se si desidera semplicemente ricavare informazioni concettuali sull'albero, escludere i surrogati. Se per alcuni casi i dati (predittore) della variabile indipendente sono incompleti e si desiderano regole che simulino l'albero, includere i surrogati. Per ulteriori informazioni, vedere l'argomento Surrogati in il capitolo 1 a pag. 16.

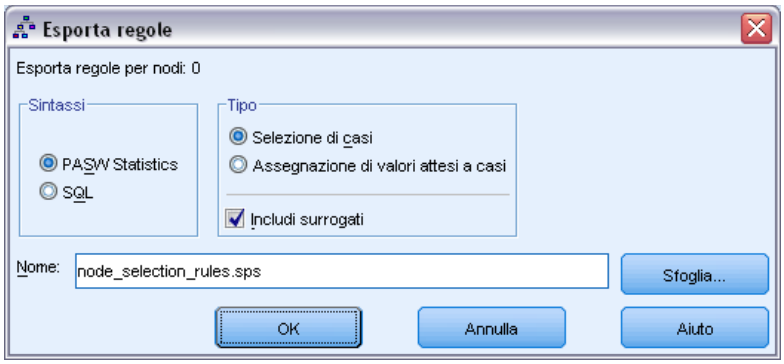
Per salvare regole di selezione o di punteggio:

▶ Selezionare i nodi nell'Editor degli alberi. Fare clic tenendo premuto il tasto Ctrl per selezionare più nodi.

▶ Dai menu, scegliere:
Regole > Esporta...

▶ Selezionare il tipo di regole desiderato e inserire un nome di file.

Figura 2-10
Finestra di dialogo Esporta regole



Nota: se si applicano le regole sotto forma di sintassi di comando a un altro file dati, questo dovrà contenere variabili con gli stessi nomi delle variabili indipendenti incluse nel modello finale, misurate nella stessa metrica e con gli stessi valori mancanti definibili dall'utente (se presenti).

Parte II: Esempi

Capitolo
3

Ipotesi sui dati e requisiti

La procedura Albero decisionale si basa sui seguenti presupposti:

- Il livello di misurazione appropriato è stato assegnato a tutte le variabili dell'analisi.
- Per le variabili dipendenti categoriali (**nominali**, **ordinali**), le etichette dei valori sono state definite per tutte le categorie da includere nell'analisi.

Verrà utilizzato il file *tree_textdata.sav* per illustrare l'importanza di entrambi questi requisiti. Il file di dati riflette lo stato predefinito dei dati letti o inseriti prima di definire gli attributi, ad esempio livello di misurazione o etichette dei valori. Per ulteriori informazioni, vedere l'argomento File di esempio in l'appendice A in *IBM SPSS Decision Trees 19*.

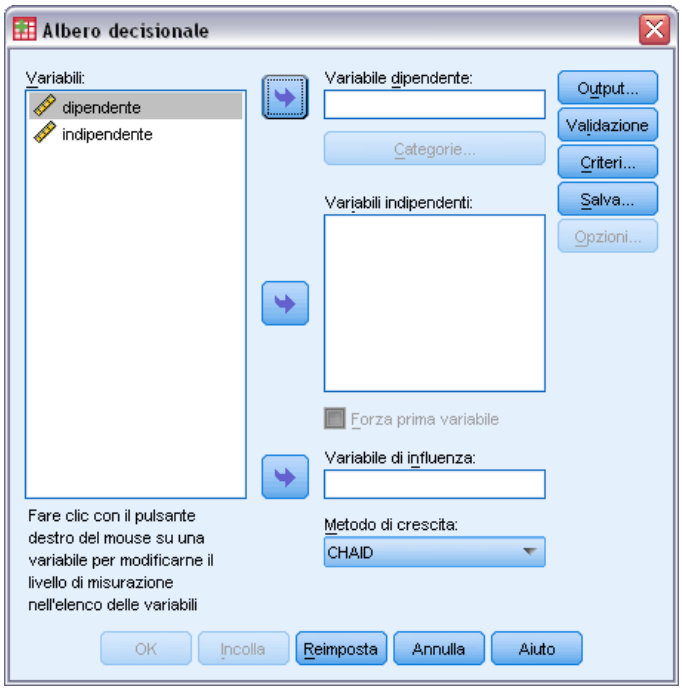
Effetti del livello di misurazione sui modelli di alberi

Entrambe le variabili di questo file di dati sono numeriche e a entrambe è stato assegnato il livello di misurazione **scala**. Tuttavia, come si vedrà in seguito, entrambe le variabili sono effettivamente variabili categoriali basate su codici numerici che indicano valori di categoria.

▶ Per eseguire un'analisi Albero decisionale, dai menu scegliere:
Analizza > Classifica > Albero...

L'icona accanto alle due variabili nell'elenco di variabili sorgenti indica che saranno considerate come variabili di scala.

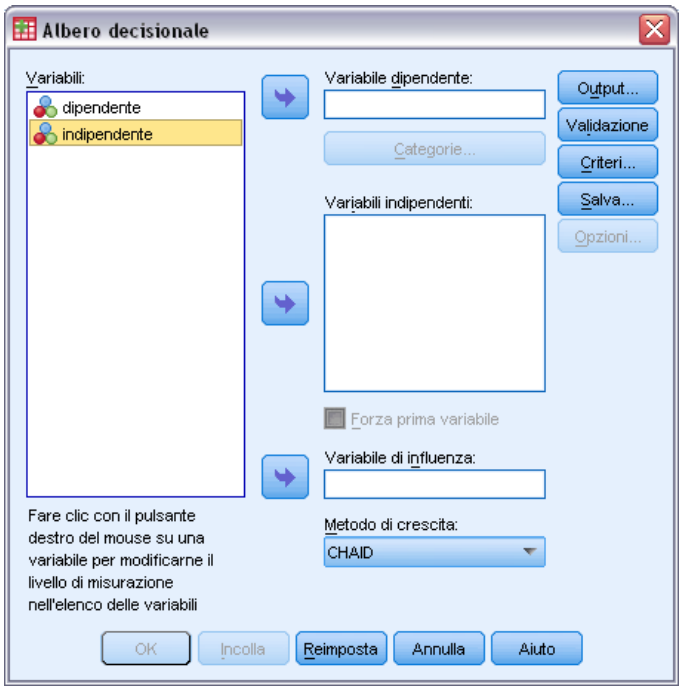
Figura 3-1
Finestra di dialogo principale Albero decisionale con due variabili di scala



- Selezionare *dipendente* come variabile dipendente.
- Selezionare *indipendente* come variabile indipendente.
- Fare clic su OK per eseguire la procedura.
- Aprire di nuovo la finestra di dialogo Albero decisionale e fare clic su Ripristina.
- Fare clic con il pulsante destro del mouse su *dipendente* nell'elenco sorgente e selezionare Nominale dal menu di scelta rapida.
- Eseguire la stessa procedura per la variabile *dipendente* nell'elenco sorgente.

Le icone accanto a ciascuna variabile indicano che saranno considerate come variabili nominali.

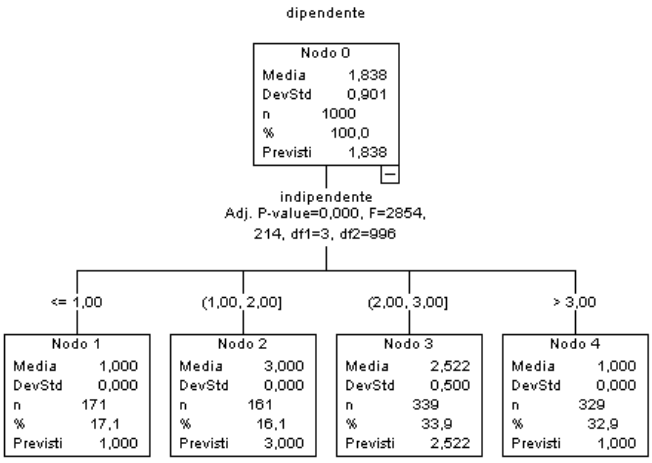
Figura 3-2
Icone nominali nell'elenco sorgente



- Selezionare *dipendente* come variabile dipendente e *indipendente* come variabile indipendente, quindi scegliere OKper ripetere la procedura.

Si procederà ora al confronto tra i due alberi. Per prima cosa verrà esaminato l'albero in cui entrambe le variabili numeriche sono considerate come variabili di scala.

Figura 3-3
Albero con entrambe le variabili considerate come variabili di scala

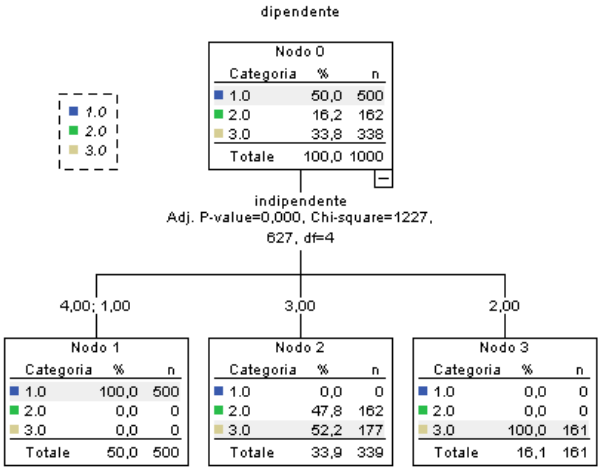


- Ciascun nodo dell'albero mostra il valore "previsto", ovvero il valore medio della variabile dipendente per quel nodo. Per una variabile che sia effettivamente categoriale, la media potrebbe non essere una statistica significativa.
- L'albero ha quattro nodi figlio, uno per ciascun valore della variabile indipendente.

Nei modelli di albero spesso i nodi simili vengono uniti, ma per una variabile di scala è possibile unire solo valori consecutivi. Nell'esempio, non sono presenti valori consecutivi considerati sufficientemente simili per procedere all'unione di due o più nodi.

L'albero in cui entrambe le variabili sono considerate come nominali è diverso per alcuni aspetti.

Figura 3-4
Albero con entrambe le variabili considerate come nominali



- Anziché un valore atteso, ciascun nodo include una tabella di frequenza che mostra il numero di casi (conteggio e percentuale) per ciascuna categoria della variabile dipendente.
- La categoria "prevista"—quella con il conteggio maggiore in ciascun nodo—è evidenziata. Ad esempio, la categoria prevista per il nodo 2 è la categoria 3. .
- Anziché quattro, sono presenti solo tre nodi figlio, con due valori della variabile indipendente uniti in un unico nodo.

I due valori indipendenti uniti nello stesso nodo sono l'1 e il 4. Poiché, per definizione, i valori nominali non hanno ordine intrinseco, è consentita l'unione di valori non consecutivi.

Assegnazione permanente del livello di misurazione

Quando si modifica il livello di misurazione per una variabile nella finestra di dialogo Albero decisionale, la modifica è solo temporanea e non viene salvata nel file di dati. Inoltre, non è sempre possibile conoscere il livello di misurazione corretto per tutte le variabili.

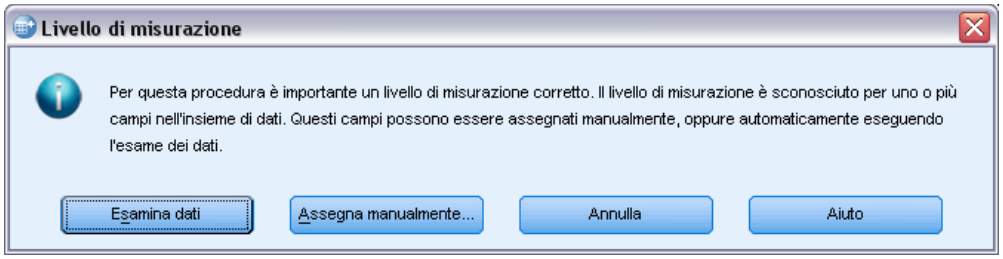
Definisci proprietà variabili può consentire di determinare il livello di misurazione corretto per ciascuna variabile e di modificare in modo permanente il livello di misurazione assegnato. Per utilizzare Definisci proprietà variabili:

- Dai menu, scegliere:
  Dati > Definisci proprietà variabili...

Variabili 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 3-5
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.

Effetti delle etichette dei valori sui modelli di alberi

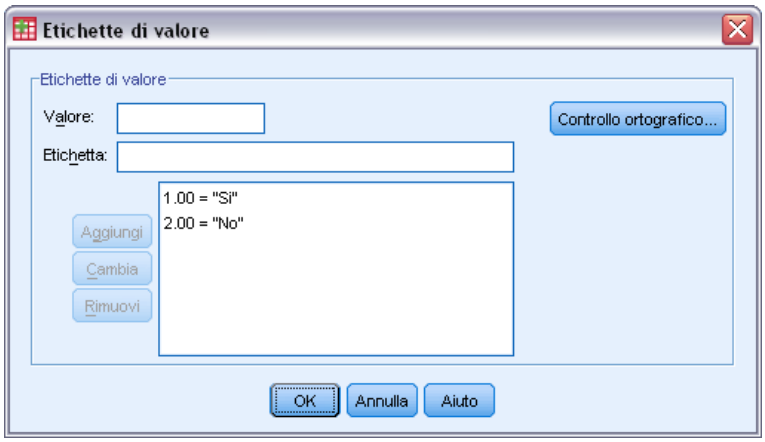
L'interfaccia della finestra di dialogo Albero decisionale presuppone che per *tutti* o per *nessuno* dei valori non mancanti di una variabile dipendente categoriale (nominale, ordinale) siano presenti etichette dei valori. Alcune funzioni non sono disponibili a meno che per almeno due valori non mancanti della variabile dipendente categoriale siano presenti etichette dei valori. Se per almeno due valori non mancanti sono state definite etichette dei valori, qualsiasi caso con altri valori privi di etichette sarà escluso dall'analisi.

Il file di dati originale nell'esempio corrente non include etichette dei valori definite; quando la variabile dipendente è considerata come nominale, il modello di albero utilizza tutti i valori non mancanti nell'analisi. In questo esempio, tali valori sono 1, 2 e 3.

Cosa avviene quando si definiscono etichette dei valori solo per alcuni valori della variabile dipendente?

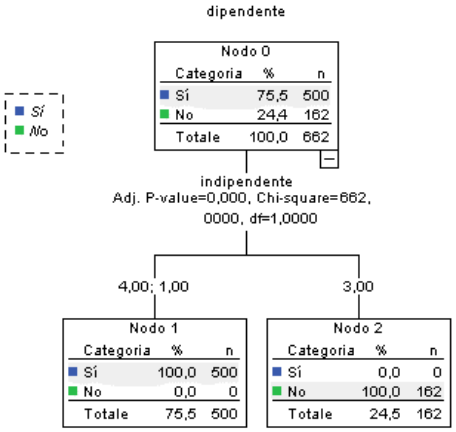
- Fare clic sulla scheda Visualizzazione variabili nella finestra dell'Editor dei dati.
- Fare clic sulla cella Valori per la variabile *dipendente*.

Figura 3-6
Definizione delle etichette dei valori per una variabile dipendente



- Per prima cosa specificare 1 come Valore e Sì come Etichette dei valori, quindi fare clic su Aggiungi.
- Quindi inserire 2 come Valore e No come Etichette dei valori, quindi fare clic su Aggiungi.
- Fare clic su OK.
- Aprire di nuovo la finestra di dialogo Albero decisionale. Nella finestra di dialogo dovrebbe essere ancora selezionato *dipendente* come variabile dipendente, con livello di misurazione nominale.
- Scegliere OK per eseguire di nuovo la procedura.

Figura 3-7
Albero per la variabile dipendente nominale con etichette dei valori parziali



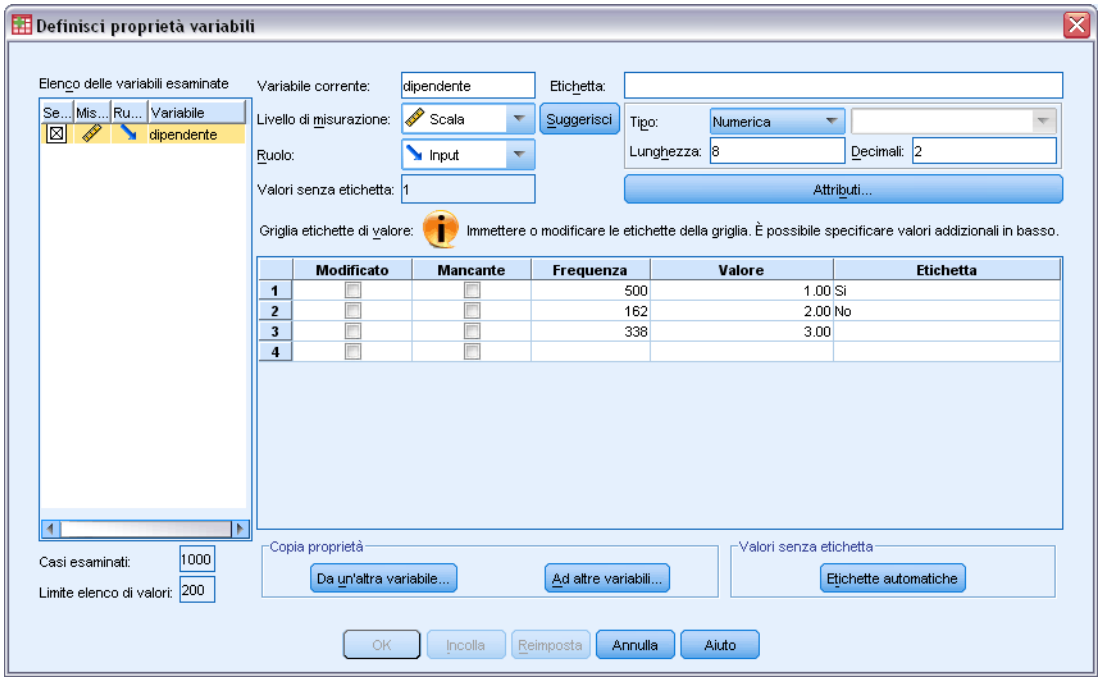
Attualmente solo due valori di variabili dipendenti con etichette dei valori definite sono inclusi nel modello di albero. Tutti i casi con valore 3 per la variabile dipendente sono stati esclusi, il che potrebbe non essere subito evidente se non si ha familiarità con i dati.

Assegnazione di etichette dei valori a tutti i valori

Per evitare l'omissione accidentale di valori categoriali validi dall'analisi, utilizzare Definisci proprietà variabili per assegnare etichette dei valori a tutti i valori di variabili dipendenti individuati nei dati.

Quando le informazioni sul dizionario dati per la variabile *nome* vengono visualizzate nella finestra di dialogo Definisci proprietà variabili, è possibile vedere che, sebbene ci siano oltre 300 casi con un valore 3 per la variabile, nessuna etichetta dei valori è stata definita per il valore.

Figura 3-8
Variabile con etichette dei valori parziali nella finestra di dialogo Definisci proprietà variabili



Capitolo
4

Utilizzo degli alberi decisionali per la valutazione del rischio di credito

Una banca conserva un database di informazioni cronologiche relative ai clienti che hanno acceso prestiti presso la banca, che include dati sul rimborso o meno dei prestiti stessi. Utilizzando il modello ad albero, è necessario analizzare le caratteristiche dei due gruppi di clienti e creare modelli per prevedere la probabilità che i richiedenti di un prestito siano inadempimenti rispetto al rimborso.

I dati di credito vengono memorizzati in *tree_credit.sav*. Per ulteriori informazioni, vedere l'argomento File di esempio in l'appendice A in *IBM SPSS Decision Trees 19*.

Creazione del modello

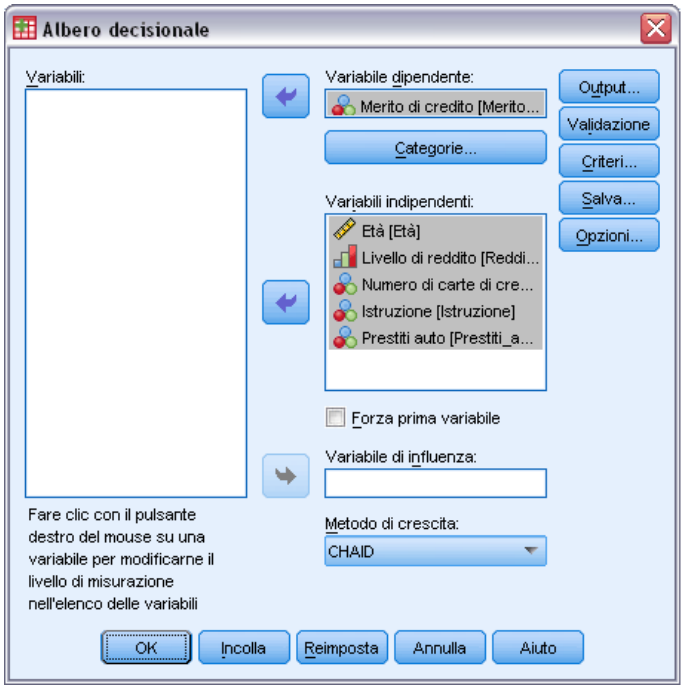
La procedura Albero decisionale offre vari metodi diversi per la creazione dei modelli ad albero. In questo esempio, viene utilizzato il metodo predefinito:

CHAID. Acronimo di Chi-squared Automatic Interaction Detection. Per ogni passaggio, CHAID scegliere la variabile (predittore) indipendente con la più forte interazione con la variabile dipendente. Le categorie di ogni predittore sono unite se non sono diverse in modo rilevante dalla variabile dipendente.

Creazione del modello di albero CHAID

- Per eseguire un'analisi Albero decisionale, dai menu scegliere:
  Analizza > Classifica > Albero...

Figura 4-1
Finestra di dialogo Albero decisionale



- Selezionare *Valutazione credito* come variabile dipendente.
- Selezionare tutte le variabili rimanenti come indipendenti. (la procedura escluderà automaticamente eventuali variabili che non contribuiscano in modo significativo al modello finale).

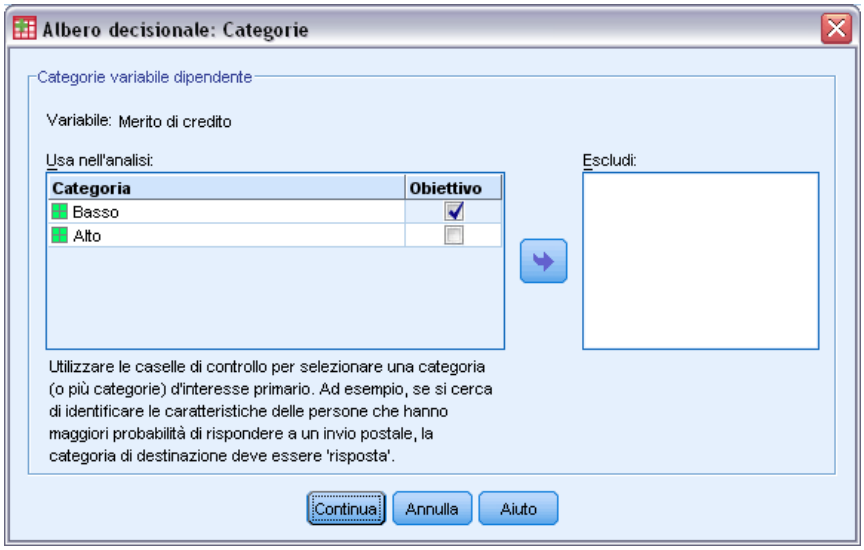
  A questo punto, è possibile eseguire la procedura e generare un modello di albero di base. Tuttavia, nell'esempio si procederà alla selezione di alcuni output aggiuntivi e all'esecuzione di alcune leggere correzioni ai criteri utilizzati per la generazione del modello.

Selezione delle categorie obbiettivo

- Fare clic sul pulsante Categorie a destra direttamente sotto la variabile dipendente selezionata.

Verrà aperta la finestra di dialogo Categorie, dove è possibile specificare le categorie obbiettivo delle variabili dipendenti di interesse. Le categorie obbiettivo non influenzano direttamente il modello di albero, ma alcuni output e opzioni sono disponibili sono tali categorie sono state selezionate.

Figura 4-2
Finestra di dialogo Categorie



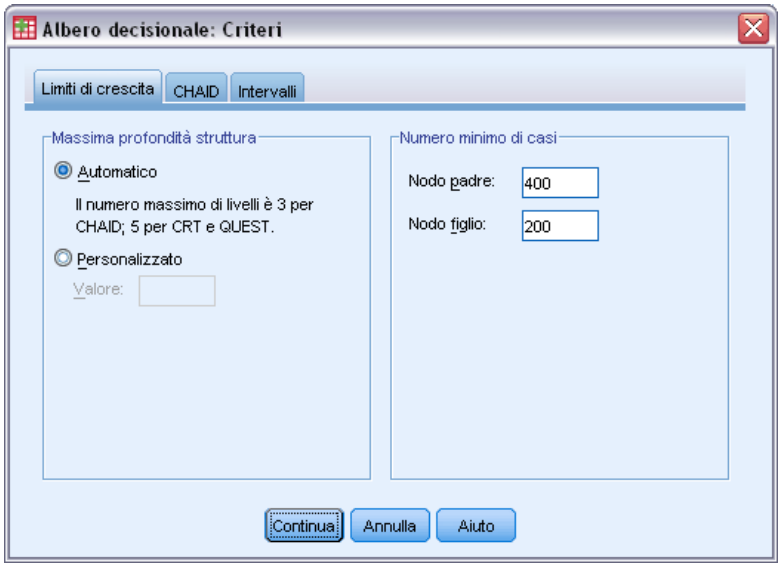
- Selezionare la casella di controllo Obbiettivo per la categoria *Negativo*. I clienti con valutazione del credito negativa (inadempienti rispetto al rimborso di un prestito) saranno considerati come categoria obbiettivo di interesse.
- Fare clic su Continua.

Specificazione dei criteri di espansione dell'albero

Nell'esempio corrente si vuole mantenere l'albero piuttosto semplice; per questo si limiterà l'espansione dell'albero aumentando il numero minimo di casi per i nodi genitore e i nodi figlio.

- Nella finestra di dialogo principale Albero decisionale fare clic su Criteri.

Figura 4-3
Finestra di dialogo Criteri, scheda Limiti di crescita



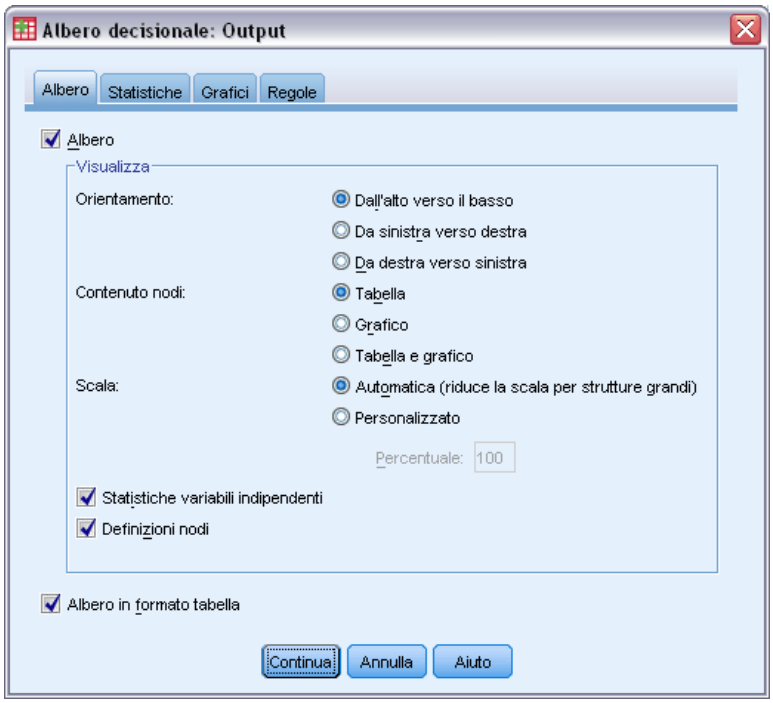
- Nel gruppo Numero minimo di casi, digitare 400 per Nodo genitore e 200 per Nodo figlio.
- Fare clic su Continua.

Selezione di output aggiuntivo

- Nella finestra di dialogo principale Albero decisionale fare clic su Output.

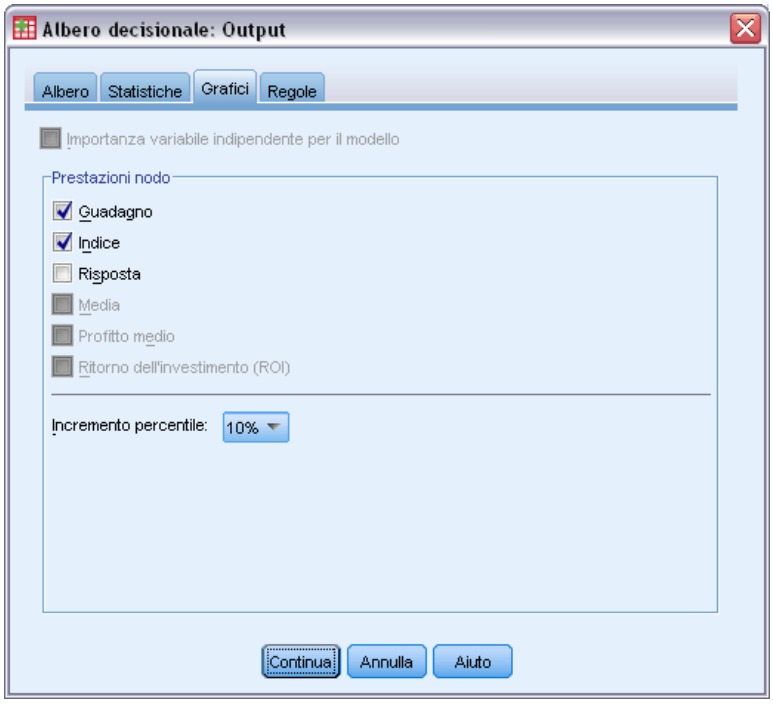
Si aprirà una finestra di dialogo a scheda, in cui sarà possibile selezionare vari tipi di output aggiuntivo.

Figura 4-4
Finestra di dialogo Output, scheda Albero



- Nella scheda Albero, selezionare Albero in formato tabella.
- Quindi fare clic sulla scheda Grafici.

Figura 4-5
Finestra di dialogo Output, scheda Grafici



▶ Selezionare Guadagno e Indice.

Nota: Questi grafici richiedono una categoria obbiettivo per la variabile dipendente. Nell'esempio corrente, la scheda Grafici non è accessibile prima di aver specificato una o più categorie obbiettivo.

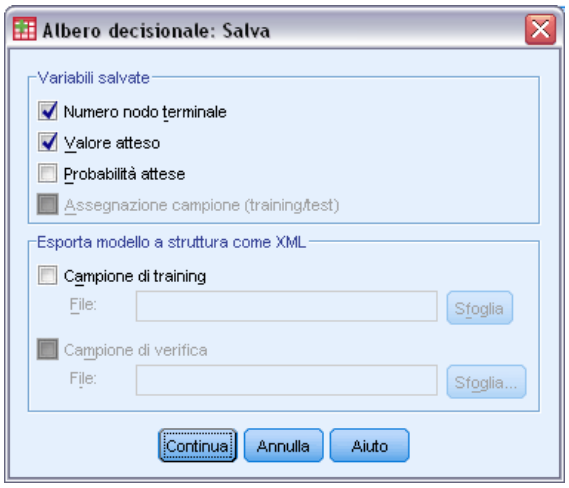
▶ Fare clic su Continua.

Salvataggio di valori attesi

È possibile salvare le variabili che contengono informazioni sulle previsioni dei modelli. Ad esempio, è possibile salvare la valutazione del credito prevista per ciascun caso e quindi confrontare tali previsioni con le valutazioni effettive.

▶ Nella finestra di dialogo principale Albero decisionale fare clic su Salva.

Figura 4-6
Salva



- Selezionare Numero nodo terminale, Valore atteso e Probabilità previste.
- Fare clic su Continua.
- Nella finestra di dialogo principale Albero decisionale fare clic su OK per eseguire la procedura.

Valutazione del modello

Per questo esempio i risultati del modello includono:

- Tabelle che forniscono informazioni sul modello.
- Diagramma ad albero.
- Tabelle che forniscono un'indicazione sull'attendibilità del modello.
- Variabili di previsione del modello aggiunte al file di dati attivo.

Tabella Riepilogo del modello

Figura 4-7
Riepilogo modello

Specifiche	Metodo di crescita	CHAID
	Variabile dipendente	Merito di credito
	Variabili indipendenti	Età, Livello di reddito, Numero di carte di credito, Istruzione, Prestiti auto
	Convalida	NONE
	Massima profondità struttura	3
	Numero minimo di casi nel nodo padre	400
	Numero minimo di casi nel nodo figlio	200
Risultati	Variabili indipendenti incluse	Livello di reddito, Numero di carte di credito, Età
	Numero di nodi	10
	Numero di nodi terminali	6
	Profondità	3

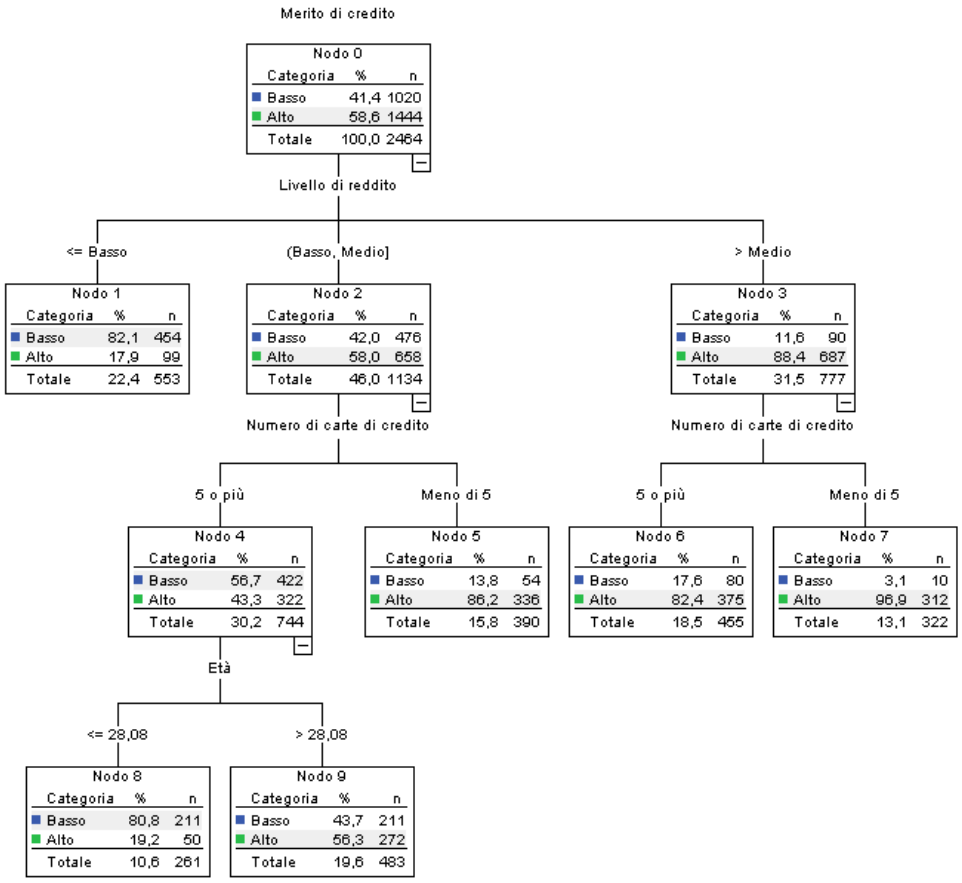
La tabella di riepilogo del modello offre informazioni molto dettagliate sulle specifiche utilizzate per creare il modello e sul modello risultante.

- La sezione Specifiche offre informazioni sulle impostazioni utilizzate per generare il modello di albero, comprese le variabili utilizzate nell'analisi.
- La sezione Risultati visualizza informazioni sul numero di nodi terminali e totali, sulla profondità dell'albero (numero di livello sotto il nodo radice e sulle variabili indipendenti incluse nel modello finale.

Sono state specificate cinque variabili indipendenti, ma solo tre sono state incluse nel modello finale. Le variabili per *istruzione* e numero di *finanziamenti auto* correnti non contribuiscono in modo significativo al modello, quindi sono state escluse automaticamente dal modello finale.

Diagramma ad albero

Figura 4-8
Diagramma ad albero per il modello di valutazione del credito



Il diagramma ad albero è una rappresentazione grafica del modello dell'albero. Questo diagramma ad albero mostra che:

- Utilizzando il metodo CHAID, *livello di reddito* costituisce il migliore predittore della *valutazione credito*.
- Per la categoria basso reddito, *livello di reddito* costituisce l'unico predittore significativo della *valutazione credito*. Tra i clienti della banca in questa categoria, l'82% si è reso inadempiente nel pagamento di prestiti. Poiché al di sotto di esso non ci sono nodi figlio, quello corrente è considerato un nodo **finale**.
- Per le categorie reddito medio e alto, il migliore predittore successivo è *numero di carte di credito*.
- Per i clienti con reddito medio e cinque o più carte di credito, il modello include un ulteriore predittore: *età*. Oltre l'80% di questi clienti fino a 28 anni di età compresi hanno una valutazione del credito negativa, mentre leggermente meno della metà di quelli di età superiore a 28 anni hanno una valutazione negativa.

È possibile utilizzare l'Editor degli alberi per nascondere e visualizzare rami selezionati, modificare colori e caratteri e selezionare sottoinsiemi basati su nodi selezionati. Per ulteriori informazioni, vedere l'argomento Selezione di casi nei nodi a pag. 74.

Tabella albero

Figura 4-9
Tabella albero per la valutazione del credito

	Basso		Alto		Totale			
Nodo	N	Percentuale	N	Percentuale	N	Percentuale	Categoria prevista	Nodo padre
0	1020	41,4%	1444	58,6%	2464	100,0%	Alto	
1	454	82,1%	99	17,9%	553	22,4%	Basso	0
2	476	42,0%	658	58,0%	1134	46,0%	Alto	0
3	90	11,6%	687	88,4%	777	31,5%	Alto	0
4	422	56,7%	322	43,3%	744	30,2%	Basso	2
5	54	13,8%	336	86,2%	390	15,8%	Alto	2
6	80	17,6%	375	82,4%	455	18,5%	Alto	3
7	10	3,1%	312	96,9%	322	13,1%	Alto	3
8	211	80,8%	50	19,2%	261	10,6%	Basso	4
9	211	43,7%	272	56,3%	483	19,6%	Alto	4

La tabella albero, come suggerisce il nome, offre la maggioranza delle informazioni essenziali incluse nel diagramma ad albero, in formato tabella. Per ciascun nodo, la tabella visualizza:

- Il numero e la percentuale di casi in ogni categoria della variabile dipendente.
- La categoria prevista per la variabile dipendente. In questo esempio, la categoria prevista è la categoria *valutazione credito*, con più del 50% di casi in quel nodo, poiché ci sono solo due possibili valutazioni creditizie.
- Il nodo genitore per ciascun nodo dell'albero. Si noti che il nodo 1—nodo del livello di reddito basso—non è il nodo genitore di nessun nodo. Essendo un nodo terminale, non ha nodi figlio.

Figura 4-10
Tabella albero per la valutazione del credito (continua)

Primary Independent Variable				
Variabile	Correzione per confronti multipli[a]	Chi-quadrato	df	Valori divisione
Livello di reddito	,000	662,457	2	<= Basso
Livello di reddito	,000	662,457	2	(Basso, Medio]
Livello di reddito	,000	662,457	2	> Medio
Numero di carte di credito	,000	193,113	1	5 o più
Numero di carte di credito	,000	193,113	1	Meno di 5
Numero di carte di credito	,000	38,587	1	5 o più
Numero di carte di credito	,000	38,587	1	Meno di 5
Età	,000	95,299	1	<= 28,07920581899067 6
Età	,000	95,299	1	> 28,07920581899067 6

- La variabile indipendente utilizzata per dividere il nodo.
- Il valore chi-quadrato (poiché l'albero è stato generato con il metodo CHAID), i gradi di libertà (*df*) e il livello di significatività (*Sig.*) per la divisione. A livello pratico, probabilmente si sarà interessati solo al livello di significatività, che è inferiore a 0,0001 per tutte le divisioni nel modello corrente.
- I valore o i valori della variabile indipendente per il nodo.

Nota: per variabili indipendenti ordinali e di scala, è possibile che siano visualizzati intervalli nell'albero e nella tabella albero espressi nella forma generica *(valore1, valore2]*, che essenzialmente significa "maggiore del valore1 e minore o uguale al valore2". Nell'esempio, il livello di reddito può avere solo tre valori—*Basso*, *Medio* e *Alto*—dove *(Basso, Medio]* significa semplicemente *Medio*. Analogamente, *>Medio* significa *Alto*.

Guadagni per i nodi

Figura 4-11
Guadagni per i nodi

	Nodo		Guadagno			
Nodo	N	Percentuale	N	Percentuale	Risposta	Indice
1	553	22,4%	454	44,5%	82,1%	198,3%
8	261	10,6%	211	20,7%	80,8%	195,3%
9	483	19,6%	211	20,7%	43,7%	105,5%
6	455	18,5%	80	7,8%	17,6%	42,5%
5	390	15,8%	54	5,3%	13,8%	33,4%
7	322	13,1%	10	1,0%	3,1%	7,5%

Metodo di crescita: CHAID
Variabile dipendente: Merito di credito

La tabella guadagni per i nodi fornisce un riepilogo delle informazioni relative ai nodi terminali nel modello.

- Solo i nodi terminali —in corrispondenza dei quali l'espansione dell'albero termina—sono elencati nella tabella. Spesso si sarà interessati solo ai nodi terminali, che rappresentano le previsioni di classificazione migliori per il modello.
- Poiché i valori di guadagno forniscono informazioni relative alle categorie obbiettivo, la tabella è disponibile solo se sono state specificate una o più tabelle obbiettivo. Nell'esempio, è presente una sola categoria obbiettivo, quindi esiste una sola tabella dei guadagni per i nodi.
- *N nodi* è il numero di casi in ogni nodo terminale; *Percentuale nodo* è la percentuale del numero totale di casi in ciascun nodo.
- *N guadagno* è il numero di casi in ogni nodo terminale nella categoria obbiettivo; *Percentuale di guadagno* è la percentuale di casi nella categoria obbiettivo rispetto al numero totale di casi nella categoria obbiettivo—nell'esempio, il numero e la percentuale di casi con una valutazione del credito negativa.
- Per le variabili dipendenti categoriali, *Risposta* è la percentuale di casi nel nodo che appartengono alla categoria obbiettivo specificata. Nell'esempio, si tratta delle stesse percentuali visualizzate per la categoria *Negativo* nel diagramma ad albero.
- Per le variabili dipendenti categoriali, *Indice* è il rapporto tra la percentuale di risposte per la categoria obbiettivo e la percentuale di risposte per l'intero campione.

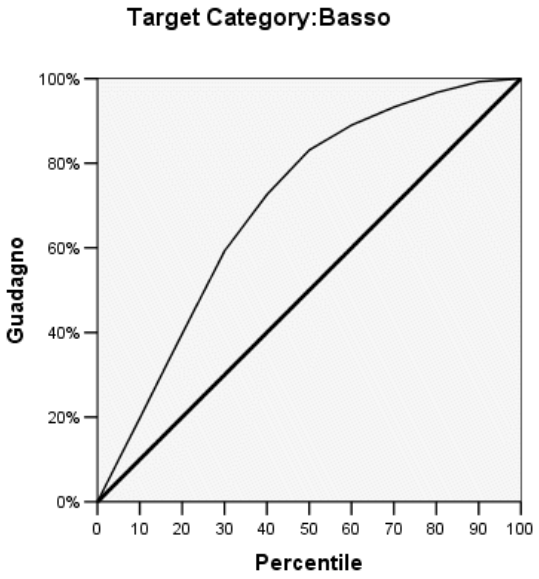
Valori indice

Il valore dell'indice è essenzialmente un'indicazione della misura in cui la percentuale della categoria obbiettivo *osservata* per il nodo differisce dalla percentuale *prevista* per la categoria obbiettivo. La percentuale della categoria obbiettivo nel nodo radice rappresenta la percentuale prevista prima di considerare gli effetti di eventuali variabili indipendenti.

Un valore dell'indice del 100% significa che ci sono più casi nella categoria obbiettivo rispetto alla percentuale globale nella categoria obbiettivo. Al contrario, un valore dell'indice minore del 100% significa che ci sono meno casi nella categoria obbiettivo rispetto alla percentuale.

Grafico Guadagni

Figura 4-12
Grafico dei guadagni per la categoria obbiettivo relativa alla valutazione del credito negativa

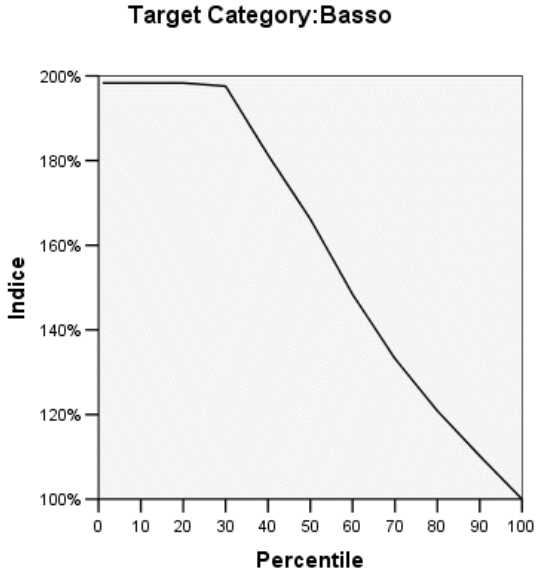


Questo grafico indica che il modello è piuttosto attendibile.

I grafici dei guadagni cumulati partono sempre dallo 0% e terminano al 100%, muovendosi da un estremo all'altro. Per un modello attendibile, il grafico dei guadagni aumenterà ripidamente verso il 100% per poi scendere. Un modello che non offre informazioni seguirà la linea di riferimento della diagonale.

Grafico indice

Figura 4-13
Grafico degli indici per la categoria obbiettivo relativa alla valutazione del credito negativa



Questo grafico indica che il modello è attendibile. I grafici degli indici cumulativi tendono a iniziare sopra il 100% e a scendere gradualmente fino a raggiungere il 100%.

In un modello attendibile, il valore dell'indice inizia molto al di sopra del 100%, rimane stabile mentre ci si sposta e quindi scende nettamente verso il 100%. Per un modello che non offre informazioni, la linea si sovrapporrà al 100% in tutto il grafico.

Stima del rischio e classificazione

Figura 4-14
Rischio e tabelle di classificazione

Rischio

Stima	Errore standard
,275	,008

Metodo di crescita: CHAID
Variabile dipendente: Merito di credito

Classificazione

	Previsione		
Osservato	Basso	Alto	Percentuale corretta
Basso	665	355	65,2%
Alto	149	1295	89,7%
Percentuale globale	33,0%	67,0%	79,5%

Metodo di crescita: CHAID
Variabile dipendente: Merito di credito

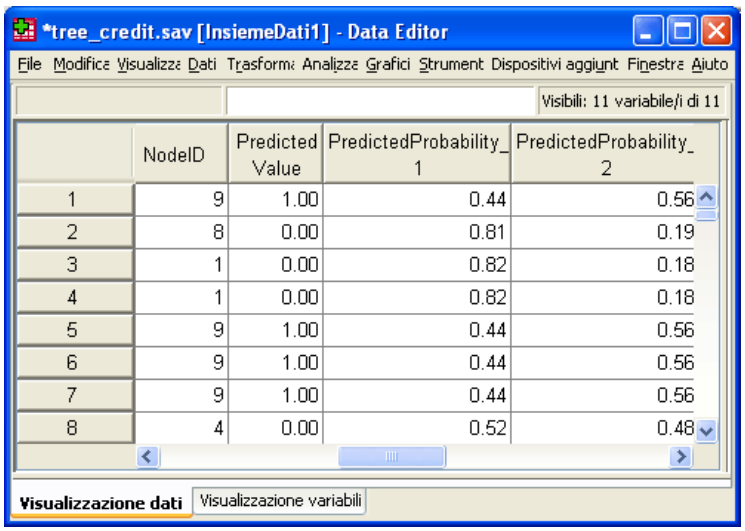
Il rischio e le tabelle di classificazione forniscono una valutazione rapida della bontà del modello.

- Una stima del rischio pari a 0,205 indica che la categoria prevista dal modello (valutazione del credito positiva o negativa) è errata nel 20,5% dei casi. Di conseguenza, il "rischio" di classificare un cliente in modo errato è di circa il 21%.
- I risultati della tabella di classificazione sono coerenti con la stima del rischio. La tabella mostra che il modello classifica correttamente circa il 79,5% dei clienti.

La tabella di classificazione, tuttavia, rivela un potenziale problema con il modello corrente: solo il 65% dei clienti con valutazione del credito negativa viene classificato come tale; in altre parole, il 35% dei clienti con valutazione negativa vengono erroneamente classificati in modo positivo.

Valori attesi

Figura 4-15
Nuove variabili per valori e probabilità attesi



	NodeID	PredictedValue	PredictedProbability_1	PredictedProbability_2
1	9	1.00	0.44	0.56
2	8	0.00	0.81	0.19
3	1	0.00	0.82	0.18
4	1	0.00	0.82	0.18
5	9	1.00	0.44	0.56
6	9	1.00	0.44	0.56
7	9	1.00	0.44	0.56
8	4	0.00	0.52	0.48

Nel file di dati attivo sono state create quattro variabili nuove:

IDNodo. Il numero del nodo terminale per ciascun caso.

ValorePrevisto. Il valore atteso della variabile dipendente per ciascun caso. Poiché la variabile dipendente è codificata 0 = *Negativo* e 1 = *Positivo*, un valore atteso pari a 0 significa si prevede che il caso abbia una valutazione del credito negativa.

ProbabilitàPrevista La probabilità che il caso appartenga a ciascuna categoria della variabile dipendente. Poiché i valori possibili sono solo due per la variabile dipendente, vengono create due variabili:

- **ProbabilitàPrevista_1.** La probabilità che il caso appartenga alla categoria di valutazione del credito negativa.
- **ProbabilitàPrevista_2.** La probabilità che il caso appartenga alla categoria di valutazione del credito positiva.

La probabilità prevista è semplicemente la proporzione di casi in ciascuna categoria della variabile dipendente per il nodo terminale che contiene ciascun caso. Ad esempio, nel nodo 1, l'82% dei casi appartengono alla categoria negativa e il 18% alla categoria positiva, con conseguenti probabilità previste, rispettivamente di 0,82 e 0,18.

Per una variabile dipendente categoriale, il valore atteso è la categoria con la maggiore proporzione di casi nel nodo terminale per ciascun caso. Ad esempio, per il primo caso, il valore atteso è 1 (valutazione positiva), poiché circa il 56% dei casi nel relativo nodo terminale hanno una valutazione positiva. Al contrario, per il secondo caso, il valore atteso è 0 (valutazione negativa), poiché circa l'81% dei casi nel relativo nodo terminale ha una valutazione negativa.

Se sono stati definiti i costi, tuttavia, la relazione tra la categoria prevista e le probabilità previste potrebbe non essere così diretta. Per ulteriori informazioni, vedere l'argomento Assegnazione dei costi ai risultati a pag. 78.

Perfezionamento del modello

In linea generale, il modello ha un tasso di corretta classificazione lievemente inferiore all'80%. Questo si riflette nella maggioranza dei nodi terminali, dove la categoria prevista —la categoria evidenziata nel nodo—è la stessa della categoria effettiva per almeno l'80% dei casi.

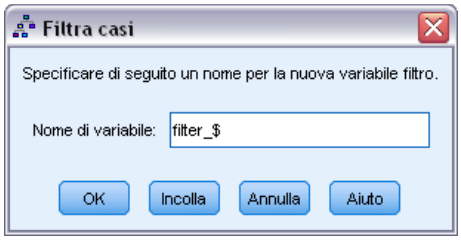
Esiste tuttavia un nodo terminale in cui i casi sono divisi in modo piuttosto uniforme tra valutazioni creditizie positive e negative. Nel nodo 9, la valutazione prevista è "positiva", ma solo il 56% dei casi nel nodo hanno effettivamente una valutazione positiva. Questo significa che a quasi la metà dei casi nel nodo (44%) verrà assegnata la categoria prevista errata. Qualora la preoccupazione principale sia l'identificazione dei rischi di credito negativi, questo nodo non risulta molto attendibile.

Selezione di casi nei nodi

Si esaminino i casi nel nodo 9 per verificare se i dati rivelano eventuali informazioni aggiuntive.

- Fare doppio clic sull'albero nel Viewer per aprire l'Editor degli alberi.
- Fare clic sul nodo 9 per selezionarlo. Per selezionare più nodi fare clic tenendo premuto Ctrl.
- Dai menu dell'Editor degli alberi, scegliere:
  Regole > Filtra casi...

Figura 4-16
Finestra di dialogo Filtra casi



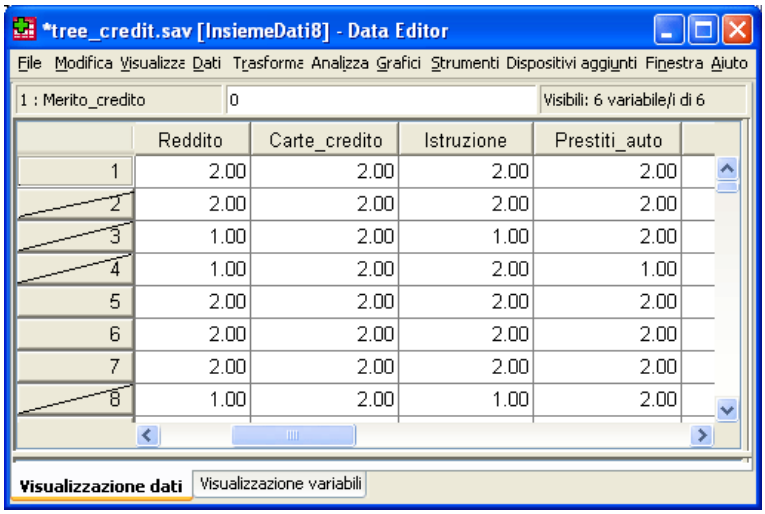
La finestra di Filtra casi consente di creare una variabile di filtro e di applicare un'impostazione di filtro in base ai valori della variabile. Il nome della variabile di filtro predefinita è *filter_$*.

- I casi dei nodi selezionati riceveranno un valore pari a 1 per la variabile di filtro.
- Tutti gli altri casi riceveranno un valore pari a 0 e saranno esclusi dalle analisi successive fino a quando non viene modificato lo stato del filtro.

Nell'esempio, questo significa che i casi non inclusi nel nodo 9 verranno filtrati per ora (ma non eliminati).

- Fare clic su OK per creare la variabile di filtro e applicare la condizione di filtro.

Figura 4-17
Casi filtrati nell'Editor dei dati



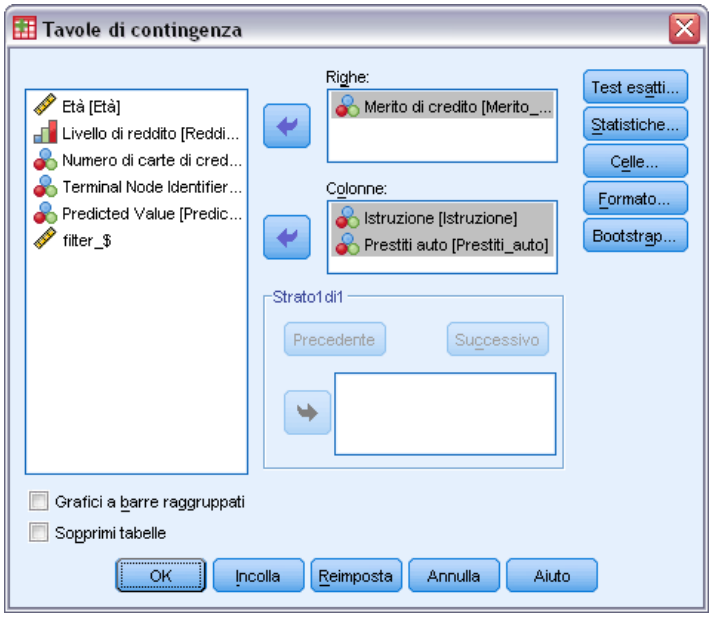
Nell'Editor dei dati, i casi filtrati sono indicati da una barra diagonale sopra il numero della riga. I casi non inclusi nel nodo 9 sono filtrati. I casi nel nodo 9 non vengono filtrati; in questo modo le analisi successive includeranno solo casi dal nodo 9.

Esame dei casi selezionati

Come primo passaggio nell'esame dei casi nel nodo 9, è possibile esaminare le variabili non utilizzate nel modello. Nell'esempio, tutte le variabili nel file di dati sono state incluse nell'analisi, ma due di esse non sono state incluse nel modello finale: *istruzione* e *finanziamenti auto*. Poiché probabilmente esiste una ragione valida per la quale la procedura ha omesso tali categorie dal modello finale, esaminarle comunque anche se è possibile che non forniscano molte informazioni.

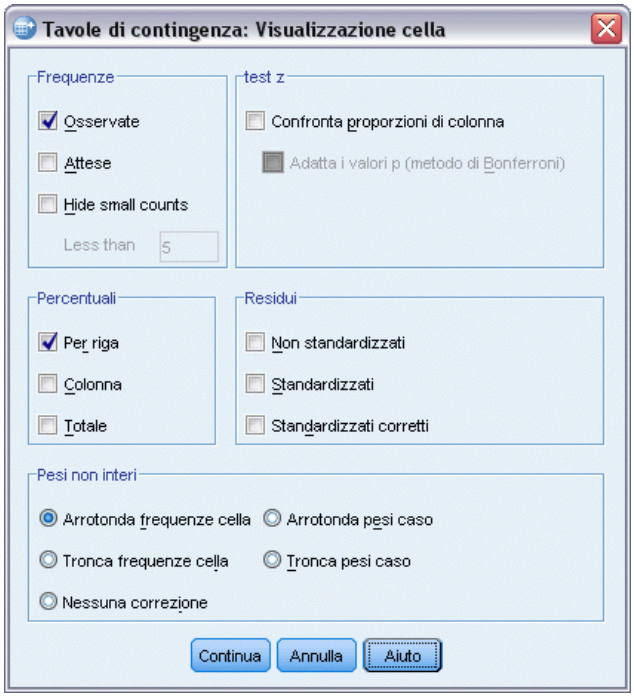
- Dai menu, scegliere:
  Analizza > Statistiche descrittive > Tavole di contingenza...

Figura 4-18
Finestra di dialogo Tavole di contingenza



- Selezionare *Valutazione credito* per la variabile riga.
- Selezionare *Istruzione* e *Finanziamenti auto* per le variabili colonna.
- Fare clic su Celle.

Figura 4-19
Finestra di dialogo Tavole di contingenza: Visualizzazione cella



- Fare clic su Riga nel gruppo Percentuali.
- Fare clic su Continua e quindi scegliere OK nella finestra di dialogo principale Tavole di contingenza per eseguire la procedura.

Esaminando le tavole di contingenza, è possibile visualizzare che per le due variabili non incluse nel modello non esiste una grande differenza tra i casi nelle categorie di valutazione del credito positiva e negativa.

Figura 4-20
Tavole di contingenza per i casi nel nodo selezionato

Tavola di contingenza Merito di credito * Istruzione

			Istruzione		
			Scuola superiore	Università	Totale
Merito di credito	Basso	Conteggio	110	101	211
		% entro Merito di credito	52,1%	47,9%	100,0%
	Alto	Conteggio	128	144	272
		% entro Merito di credito	47,1%	52,9%	100,0%
Totale		Conteggio	238	245	483
		% entro Merito di credito	49,3%	50,7%	100,0%

Tavola di contingenza Merito di credito * Prestiti auto

			Prestiti auto		
			Nessuno o 1	2 o più	Totale
Merito di credito	Basso	Conteggio	18	193	211
		% entro Merito di credito	8,5%	91,5%	100,0%
	Alto	Conteggio	39	233	272
		% entro Merito di credito	14,3%	85,7%	100,0%
Totale		Conteggio	57	426	483
		% entro Merito di credito	11,8%	88,2%	100,0%

- Per *istruzione*, poco più della metà dei casi con una valutazione del credito negativa hanno completato la scuola superiore, mentre poco più della metà dei casi con una valutazione del credito positiva hanno un'istruzione universitaria—ma questa differenza non è significativa dal punto di vista statistico.
- Per *finanziamenti auto*, la percentuale di casi di credito positivo con un solo o nessun finanziamento auto è maggiore della percentuale corrispondente per i casi di credito negativo, ma l'ampia maggioranza dei casi in entrambi in gruppi è titolare di due o più finanziamenti.

In questo modo, sebbene sia possibile avere un'idea del motivo dell'esclusione di queste variabili nel modello finale, non ci sono informazioni su come ottenere una migliore previsione per il nodo 9. Se sono presenti altre variabili non specificate per l'analisi, è possibile che si voglia esaminarne alcune prima di procedere.

Assegnazione dei costi ai risultati

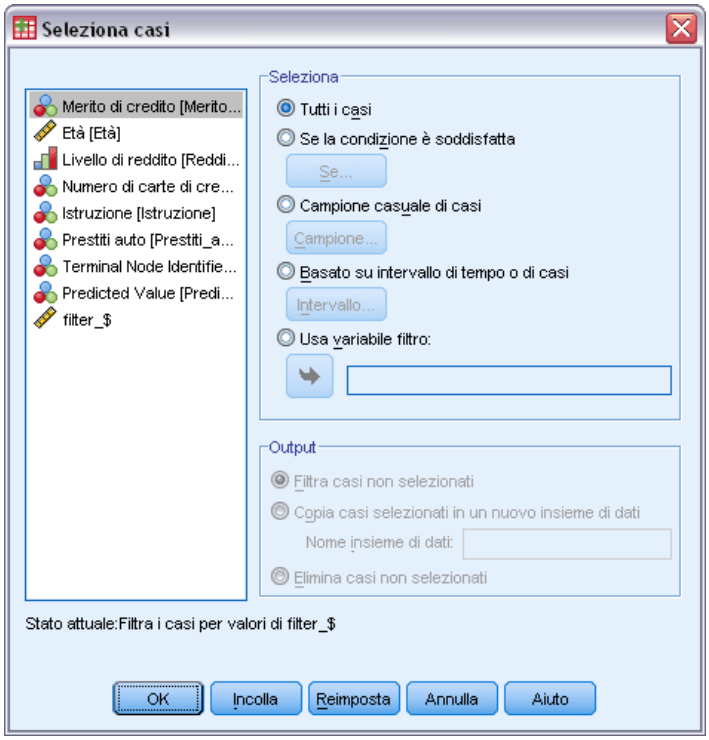
Come già indicato, a parte il fatto che quasi la metà dei casi nel nodo 9 appartengono a ciascuna categoria di valutazione del credito, il fatto che la categoria prevista sia "positiva" rappresenta un problema se l'obiettivo principale è creare un modello che identifichi correttamente i rischi di credito negativi. Sebbene possa non essere possibile migliorare le prestazioni del nodo 9, resta possibile perfezionare il modello per migliorare il tasso di corretta classificazione per i casi di valutazione negativa, anche se questo determinerà anche un tasso maggiore di errata classificazione per i casi di valutazione positiva.

Per prima cosa, è necessario disabilitare la funzione di filtro dei casi, in modo che l'analisi consideri di nuovo tutti i casi.

▶ Dai menu, scegliere:
Dati > Seleziona casi...

▶ Selezionare Tutti i casi nella finestra di dialogo Seleziona casi e quindi fare clic su OK.

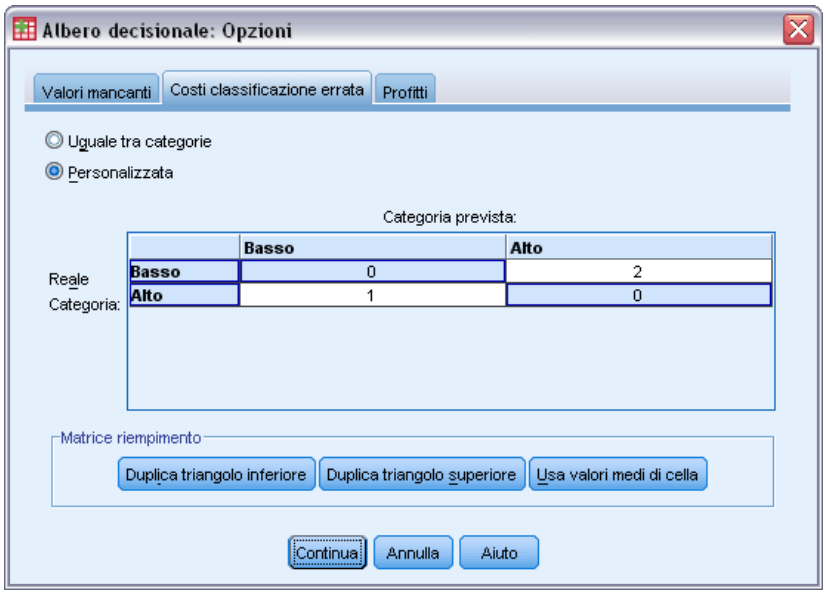
Figura 4-21
Finestra di dialogo Seleziona casi



▶ Aprire di nuovo la finestra di dialogo principale Albero decisionale e fare clic su Opzioni.

▶ Fare clic sulla scheda Costi errata classificazione.

Figura 4-22
Finestra di dialogo Opzioni, scheda Costi di errata classificazione

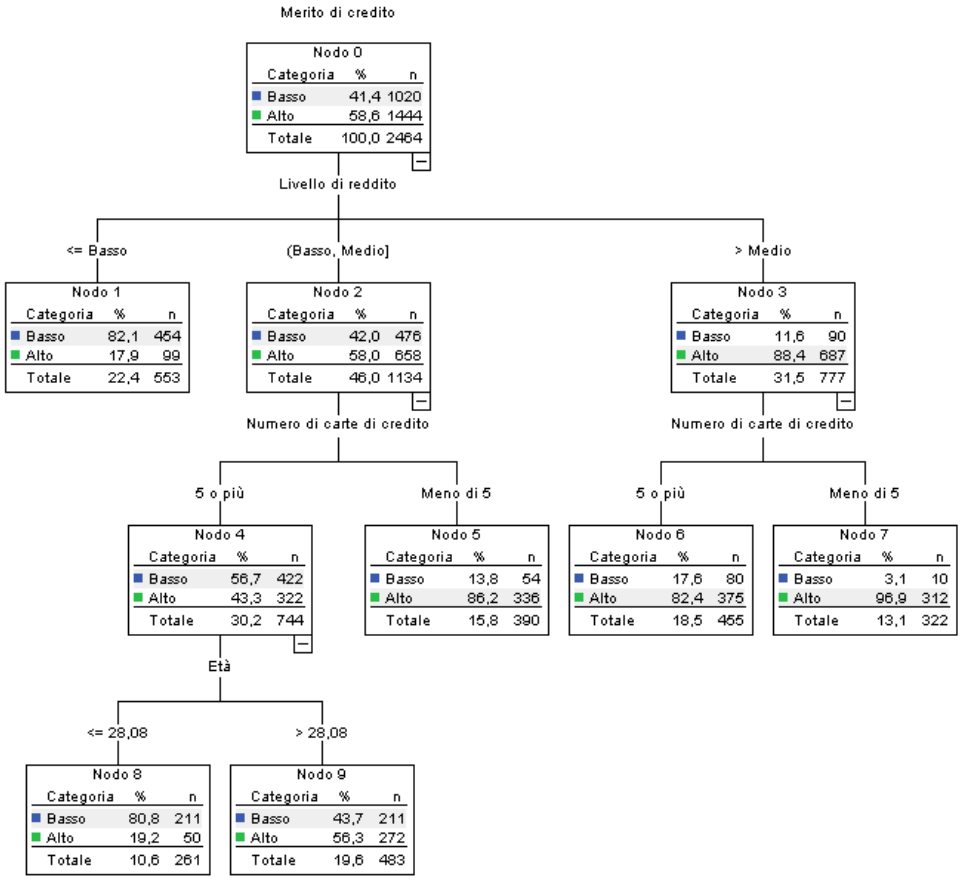


▶ Selezionare Personalizzato per Categoria effettiva *Negativo* / Categoria prevista *Positivo* specificare il valore 2.

Questo indica alla procedura che il “costo” della classificazione errata come positivo di un rischio di credito negativo è doppio rispetto alla situazione inversa.

▶ Fare clic su Continua, quindi scegliere OK nella finestra di dialogo principale per eseguire la procedura.

Figura 4-23
Modello di albero con valori di costo corretti



Al primo impatto, l'albero generato dalla procedura sembra essenzialmente uguale a quello originale. Un esame più approfondito rivela tuttavia che, sebbene la distribuzione dei casi in ciascun nodo non sia cambiata, alcune categorie previste lo sono.

Per i nodi terminali, la categoria prevista rimane la stessa in tutti i nodi eccetto uno: il nodo 9. La categoria prevista è ora *Negativo* anche se poco meno della metà dei casi appartengono alla categoria *Positivo*.

Poiché è stato specificato che il costo dell'errata classificazione dei rischi di credito negativo è superiore, a qualsiasi nodo in cui i casi siano uniformemente distribuiti tra le due categorie è ora assegnata come categoria prevista *Negativo* anche se una lieve maggioranza di casi appartiene alla categoria *Positivo*.

Questa modifica nella categoria prevista si riflette nella tabella di classificazione.

Figura 4-24
Rischio e tabelle di classificazione basate sui costi corretti

Rischio

Stima	Errore standard
,288	,011

Metodo di crescita: CHAID
Variabile dipendente: Merito di credito

Classificazione

	Previsione		
Osservato	Basso	Alto	Percentuale corretta
Basso	876	144	85,9%
Alto	421	1023	70,8%
Percentuale globale	52,6%	47,4%	77,1%

Metodo di crescita: CHAID
Variabile dipendente: Merito di credito

- Quasi l'86% dei rischi di credito negativi sono ora classificati in modo corretto, rispetto al 65% di prima.
- D'altro canto, la corretta classificazione dei rischi positivi è passata dal 90% al 71%, e la corretta classificazione globale si è ridotta dal 79,5% al 77,1%.

Si noti che anche la stima del rischio e il tasso globale di classificazione corretta non sono più coerenti tra loro. Con un tasso globale di classificazione corretta pari a 77,1%, la stima del rischio prevista sarebbe di 0,229. L'aumento del costo di errata classificazione dei casi di credito negativo, nell'esempio, ha gonfiato il valore del rischio, rendendone l'interpretazione meno diretta.

Riepilogo

È possibile usare modelli di albero per classificare i casi in gruppi identificati tramite determinate caratteristiche, ad esempio quelle associate ai clienti di una banca con record di credito positivi e negativi. Se uno specifico risultato previsto è più importante degli altri risultati possibili, è possibile perfezionare il modello per associare un costo di errata classificazione maggiore a quel risultato—ma la riduzione dei tassi di errata classificazione per un risultato comporta l'aumento dei tassi per gli altri.

Capitolo
5

Creazione di un modello di credito

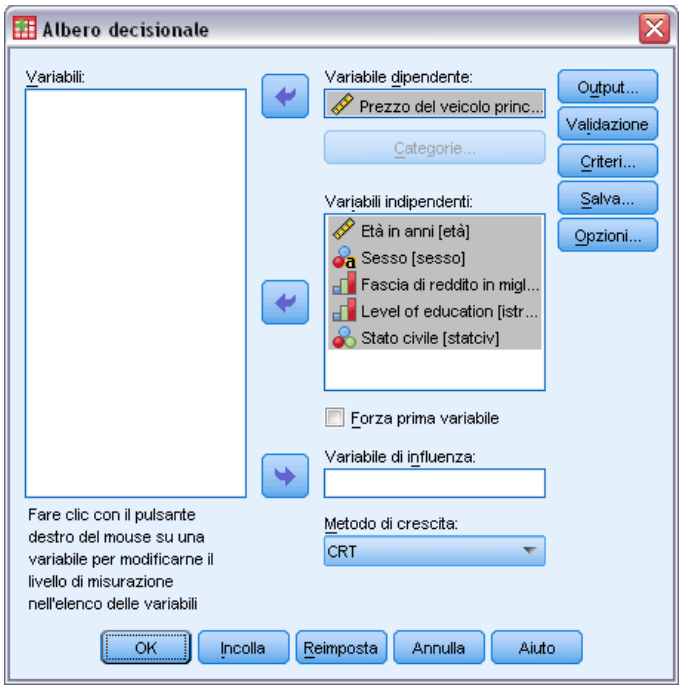
Una delle funzioni più potenti e più utili della procedura Albero decisionale è la possibilità di creare modelli successivamente applicabili ad altri file di dati per prevedere i risultati. Ad esempio, in base a un file di dati che include informazioni demografiche e sul prezzo di acquisto di un veicolo, è possibile creare un modello utilizzabile per prevedere quale sarà la spesa probabile di persone con caratteristiche demografiche simili per una nuova auto—e quindi applicare il modello ad altri file di dati in cui sono disponibili le informazioni demografiche ma non quelle relative ai precedenti acquisti di veicoli.

In questo esempio, viene utilizzato il file di dati *tree_car.sav*. Per ulteriori informazioni, vedere l'argomento File di esempio in l'appendice A in *IBM SPSS Decision Trees 19*.

Creazione del modello

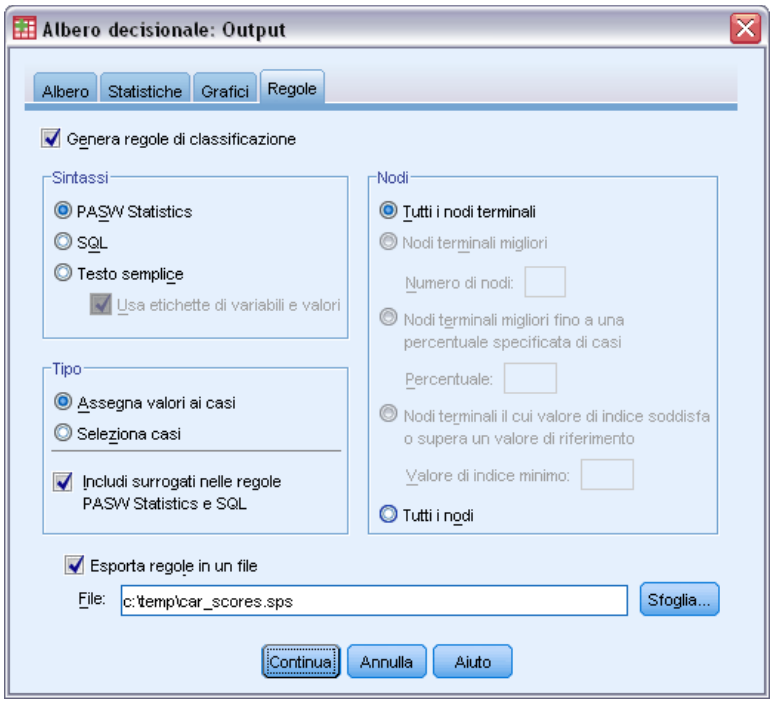
- Per eseguire un'analisi Albero decisionale, dai menu scegliere:
  Analizza > Classifica > Albero...

Figura 5-1
Finestra di dialogo Albero decisionale



- Selezionare *Prezzo del veicolo principale* come variabile dipendente.
- Selezionare tutte le variabili rimanenti come indipendenti. (la procedura escluderà automaticamente eventuali variabili che non contribuiscano in modo significativo al modello finale).
- Come metodo di espansione scegliere CRT.
- Fare clic su Output.

Figura 5-2
Finestra di dialogo Output, scheda Regole



- Fare clic sulla scheda Regole.
- Selezionare Genera regole di classificazione.
- Come Sintassi, selezionare IBM® SPSS® Statistics.
- Per Tipo, selezionare Assegna valori ai casi.
- Selezionare Esporta regola in un file e immettere un nome di file e la posizione di una directory.

  Assicurarsi di ricordare il nome del file e la posizione o annotarli. Saranno necessari in seguito. Includere un percorso di directory per essere certi di conoscere in che posizione è stato salvato il file. È possibile utilizzare il pulsante Sfoglia per passare a una posizione di directory specifica (e valida).
- Fare clic su Continua e quindi su OK per eseguire la procedura e creare il modello di albero.

Valutazione del modello

Prima di applicare il modello ad altri file di dati, è consigliabile assicurarsi che il modello funzioni correttamente con i dati originali utilizzati per la sua creazione.

Riepilogo del modello

Figura 5-3
Tabella Riepilogo del modello

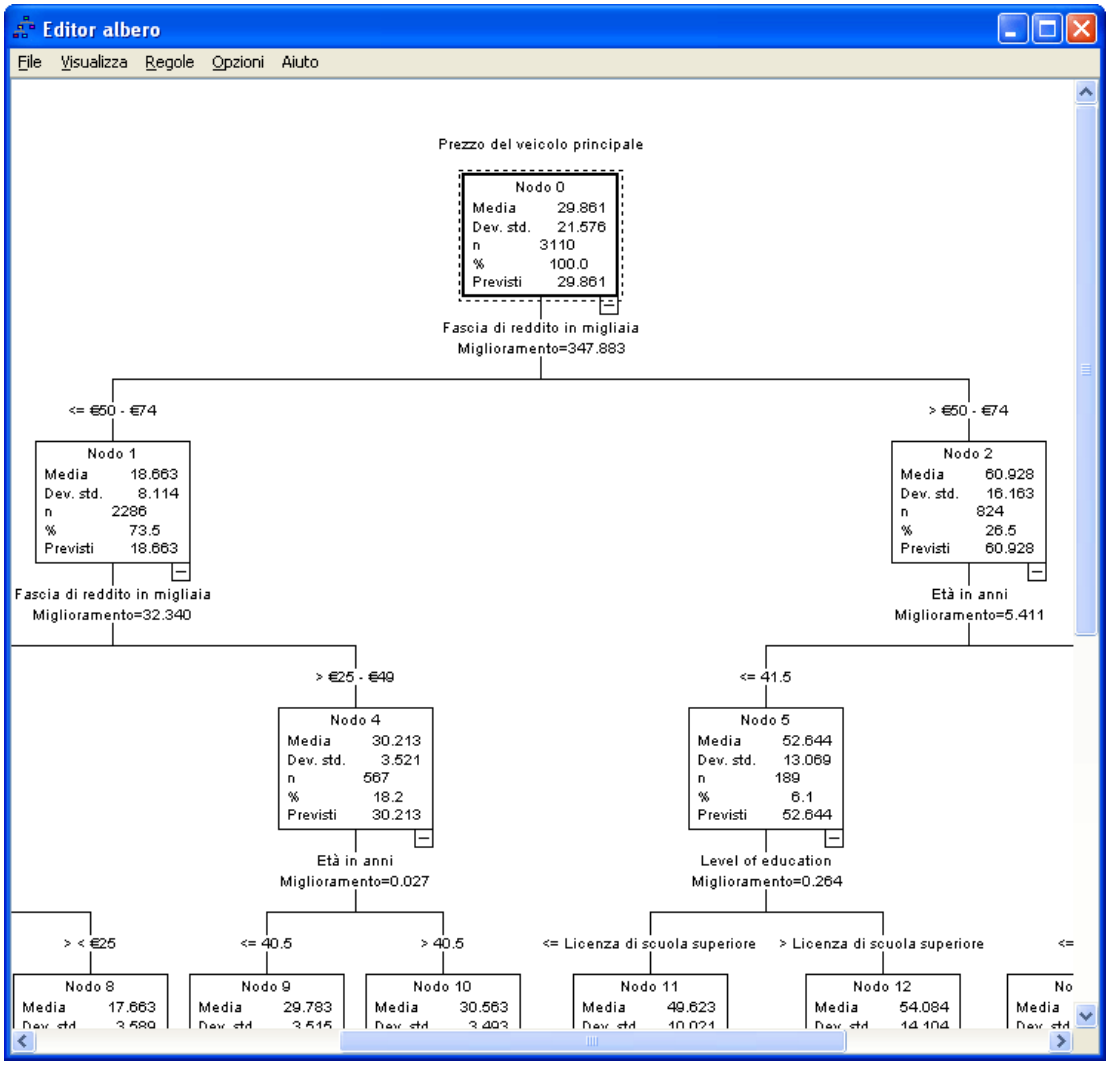
Specifiche	Metodo di crescita	CRT
	Variabile dipendente	Prezzo del veicolo principale
	Variabili indipendenti	Età in anni, Sesso, Fascia di reddito in migliaia, Level of education, Stato civile
	Convalida	NONE
	Massima profondità struttura	5
	Numero minimo di casi nel nodo padre	100
	Numero minimo di casi nel nodo figlio	50
Risultati	Variabili indipendenti incluse	Fascia di reddito in migliaia, Età in anni, Level of education
	Numero di nodi	29
	Numero di nodi terminali	15
	Profondità	5

La tabella Riepilogo del modello indica che solo tre delle variabili indipendenti selezionate hanno contribuito in modo sufficientemente significativo per essere incluse nel modello finale: *reddito*, *età* e *istruzione*. Si tratta di un'informazione importante se si desidera applicare il modello ad altri file di dati, poiché le variabili indipendenti utilizzate nel modello devono essere presenti in qualsiasi file di dati al quale applicare il modello.

La tabella di riepilogo indica inoltre che il modello di albero probabilmente non è particolarmente semplice, visto che include 29 nodi e 15 nodi terminali. Questo potrebbe non essere un problema se si desidera un modello affidabile che sia applicabile in modo pratico, anziché un modello semplice che sia facile da descrivere o da illustrare. Naturalmente, da un punto di vista pratico, è probabile che si desideri anche un modello che non si basi su un numero eccessivo di variabili (predittore) indipendenti. In questo caso non è un problema perché solo tre variabili indipendenti sono state incluse nel modello finale.

Diagramma del modello di albero

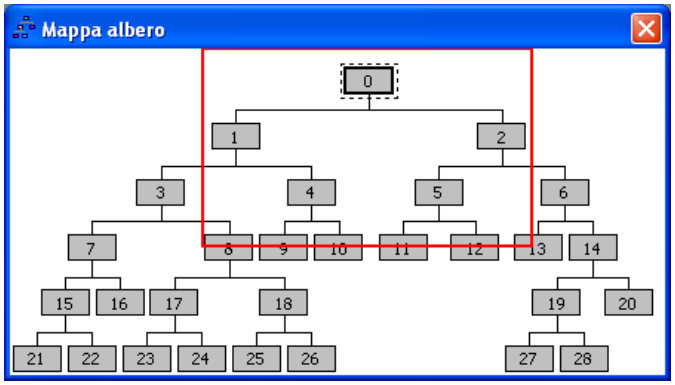
Figura 5-4
Diagramma del modello di albero albero nell'Editor degli alberi



Il diagramma del modello di albero include un numero tale di nodi da rendere difficile la visualizzazione completa del modello in dimensioni tali da poter leggere anche le informazioni sul contenuto dei nodi. È possibile utilizzare la mappa dell'albero per visualizzare l'intero albero:

- Fare doppio clic sull'albero nel Viewer per aprire l'Editor degli alberi.
- Dai menu dell'Editor degli alberi, scegliere:
  Visualizza > Mappa albero

Figura 5-5
Mappa dell'albero



- La mappa dell'albero mostra l'intero albero. È possibile modificare la dimensione della finestra della mappa per aumentare o ridurre la visualizzazione in base alle dimensioni della finestra.
- L'area evidenziata sulla mappa corrisponde alla sezione dell'albero attualmente visualizzata nell'Editor degli alberi.
- È possibile utilizzare la mappa dell'albero per spostarsi all'interno dell'albero e selezionare i nodi.

Per ulteriori informazioni, vedere l'argomento Mappa albero in il capitolo 2 a pag. 41.

Per variabili dipendenti di scala, ogni nodo mostra la deviazione standard e la media della variabile dipendente. Il nodo 0 visualizza un prezzo di acquisto del veicolo medio globale paria a circa 29,9 (in migliaia), con deviazione standard pari a 21,6.

- Il nodo 1, che rappresenta i casi con reddito inferiore a 75 (in migliaia), ha un prezzo medio pari a 18,7.
- Al contrario, il nodo 2, che rappresenta i casi con reddito uguale o superiore a 75, ha un prezzo medio pari a 60,9.

Un ulteriore esame dell'albero mostra che *età* e *istruzione* visualizzano anch'esse una relazione con il prezzo di acquisto del veicolo, ma attualmente il principale motivo di interesse è l'applicazione pratica del modello, anziché un esame dettagliato dei suoi componenti.

Stima del rischio

Figura 5-6
Tabella di rischio

Rischio

Stima	Errore standard
68,485	2,985

Metodo di crescita: CRT
Variabile dipendente: Prezzo del veicolo principale

Nessuno dei risultati esaminati finora indica se il modello è particolarmente attendibile. Un indicatore dell'attendibilità del modello è la stima del rischio. Per una variabile dipendente di scala, la stima del rischio è una misura della varianza all'interno del nodo, di per se stessa non estremamente significativa. Una varianza minore indica una maggiore attendibilità del modello, ma la varianza è relativa all'unità di misura. Se ad esempio il prezzo è stato registrato in unità anziché in migliaia, la stima del rischio sarà mille volte maggiore.

Per ottenere un'interpretazione significativa della stima del rischio con una variabile dipendente di scala è necessario esaminare quanto segue:

- la varianza totale è pari alla varianza all'interno del nodo (errore) più la varianza tra i nodi (spiegata).
- La varianza all'interno del nodo è il valore della stima del rischio: 68.485.
- La varianza totale è pari alla varianza per le variabili dipendenti prima di considerare eventuali variabili indipendenti, ovvero la varianza in corrispondenza del nodo radice.
- La deviazione standard visualizzata per il nodo radice è 21,576; di conseguenza la varianza totale è lo stesso valore elevato al quadrato: 465.524.
- La proporzione della varianza dovuta all'errore (varianza non spiegata) è 68,485/465,524 = 0,147.
- La proporzione della varianza spiegata rispetto al modello è 1–0,147 = 0,853, o 85,3%, che indica che il modello è piuttosto attendibile (l'interpretazione è simile a quella del tasso globale di corretta classificazione per una variabile dipendente categoriale).

Applicazione del modello a un altro file di dati

Avendo determinato che il modello è sufficientemente attendibile, è ora possibile applicarlo ad altri file di dati che includono variabili *età*, *reddito* e *istruzioni* analoghe, per generare una nuova variabile che rappresenti il prezzo di acquisto del veicolo previsto per ogni caso del file. Questo procedimento talvolta viene definito **assegnazione di credito**.

Alla generazione del modello, è stato specificato che le "regole" per l'assegnazione dei valori ai casi dovevano essere salvate in un file di testo con la sintassi di comando. Si utilizzeranno ora i comandi del file per generare punteggi in un altro file di dati.

▶ Aprire un file di dati *tree_score_car.sav*. Per ulteriori informazioni, vedere l'argomento File di esempio in l'appendice A in *IBM SPSS Decision Trees 19*.

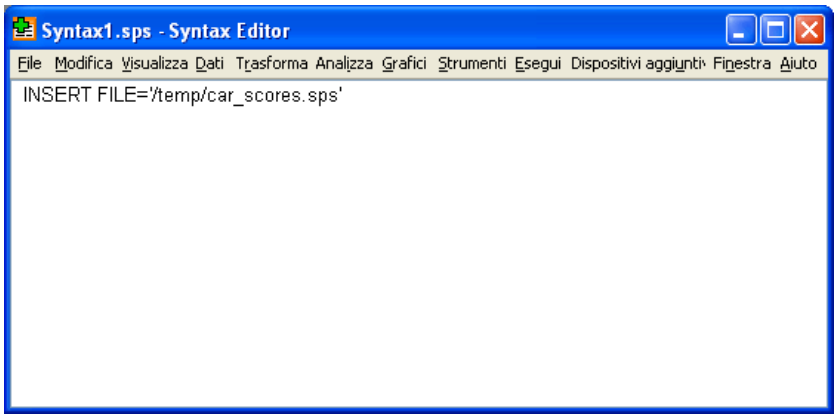
▶ Quindi, dai menu scegliere:
File > Nuovo > Sintassi

▶ Nella finestra della sintassi di comando, digitare:

```
INSERT FILE=
 '/temp/car_scores.sps'.
```

Se è stato utilizzato un nome di file o un percorso diverso, modificare la stringa di conseguenza.

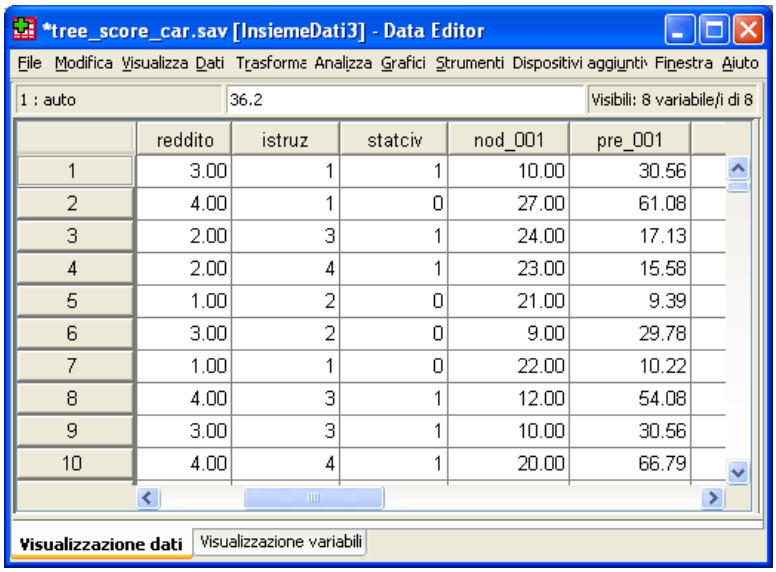
Figura 5-7
Finestra della sintassi con comando INSERT per l'esecuzione di un file di comandi



Il comando `INSERT` eseguirà i comandi nel file specificato, ovvero il file di "regole" generato alla creazione del modello.

▶ Dai menu della finestra della sintassi di comando scegliere:
Esegui > Tutto

Figura 5-8
Valori attesi aggiunti al file di dati



	reddito	istruz	statciv	nod_001	pre_001
1	3.00	1	1	10.00	30.56
2	4.00	1	0	27.00	61.08
3	2.00	3	1	24.00	17.13
4	2.00	4	1	23.00	15.58
5	1.00	2	0	21.00	9.39
6	3.00	2	0	9.00	29.78
7	1.00	1	0	22.00	10.22
8	4.00	3	1	12.00	54.08
9	3.00	3	1	10.00	30.56
10	4.00	4	1	20.00	66.79

L'operazione determina l'aggiunta di nuove variabili al file di dati:

- *nod_001* contiene il numero del nodo terminale previsto dal modello per ciascun caso.
- *pre_001* contiene il valore atteso per il prezzo di acquisto del veicolo per ciascun caso.

Poiché sono state richieste regole per l'assegnazione di valori ai nodi terminali, il numero dei possibili valori attesi è lo stesso del numero dei nodi terminali, in questo caso 15. Ad esempio, ogni caso con un numero di nodo previsto pari a 10 avrà lo stesso prezzo di acquisto del veicolo previsto: 30.56. Non per caso, si tratta del valore medio indicato per il nodo terminale 10 nel modello originale.

Sebbene il modello si applicherà generalmente a dati per i quali il valore della variabile dipendente non è noto, nell'esempio il file di dati al quale viene applicato il modello include effettivamente questa informazione—ed è possibile confrontare le previsioni del modello con i valori effettivi.

▶ Dai menu, scegliere:
Analizza > Correlazione > Bivariata...

▶ Selezionare *Prezzo del veicolo principale* e *pre_001*.

Figura 5-9
Finestra di dialogo Correlazioni bivariate

▶ Fare clic su OK per eseguire la procedura.

Figura 5-10
Correlazione tra prezzo effettivo e prezzo previsto del veicolo

		Prezzo del veicolo principale	pre_001
Prezzo del veicolo principale	Correlazione di Pearson	1	,919**
	Sig. (2-code)		,000
	N	3290	3290
pre_001	Correlazione di Pearson	,919**	1
	Sig. (2-code)	,000	
	N	3290	3290

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

La correlazione pari a 0,92 indica una correlazione positiva molto elevata tra il prezzo effettivo e previsto, ovvero una buona attendibilità del modello.

Riepilogo

La procedura Albero decisionale è utilizzabile per creare modelli successivamente applicabili ad altri file di dati per prevedere i risultati. Il file di dati di destinazione deve contenere variabili con gli stessi nomi delle variabili indipendenti incluse nel modello finale, misurate nella stessa metrica e con gli stessi valori mancanti definiti dall'utente (se presenti). Tuttavia, né la variabile dipendente né quelle indipendenti escluse dal modello finale devono essere presenti nel file di dati di destinazione.

Capitolo
6

Valori mancanti nei modelli di albero

I diversi metodi di espansione riguardano i valori mancanti per variabili (predittrici) indipendenti in vari modi diversi:

- CHAID e CHAID esaustivo considerano tutti i valori mancanti definiti dall'utente e dal sistema per ciascuna variabile indipendente come una categoria singola. Per le variabili indipendenti ordinali e di scala, questa categoria può o meno, in un secondo momento, essere unita con altre della stessa variabile indipendente, in base ai criteri di espansione.
- CRT e QUEST tentano di utilizzare i **surrogati** per le variabili (predittrici) indipendenti. Per in casi in cui il valore per la variabile è mancante, per la classificazione sono utilizzate altre variabili indipendenti con associazioni ++elevate con la variabile originale. Questi predittori alternativi sono detti surrogati.

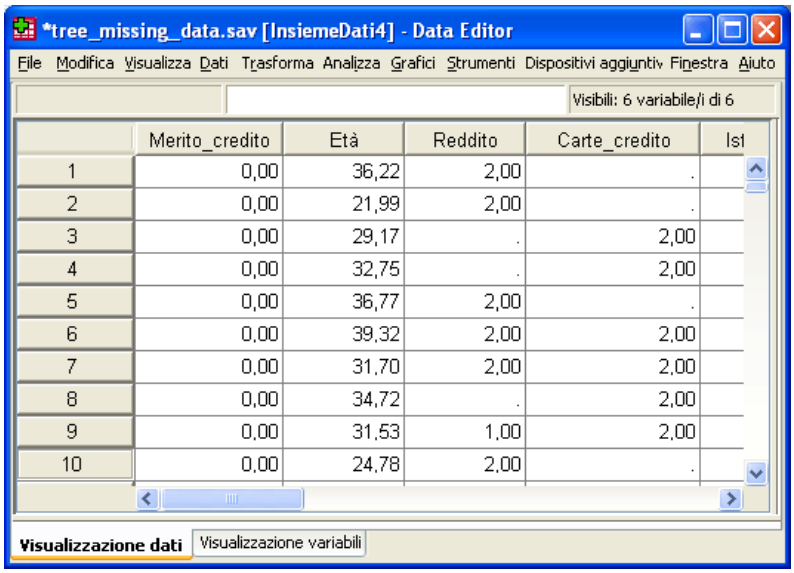
Nell'esempio è illustrata la differenza tra CHAID e CRT in presenza di valori mancanti per variabili indipendenti utilizzate nel modello.

In questo esempio, viene utilizzato il file di dati *tree_missing_data.sav*. Per ulteriori informazioni, vedere l'argomento File di esempio in l'appendice A in *IBM SPSS Decision Trees 19*.

Nota: Per variabili indipendenti nominali e variabili dipendenti nominali è possibile scegliere di considerare i valori **mancanti definiti dall'utente** come validi; tali valori verranno quindi considerati come qualsiasi altro valore non mancante. Per ulteriori informazioni, vedere l'argomento Valori mancanti in il capitolo 1 a pag. 22.

Valori mancanti con CHAID

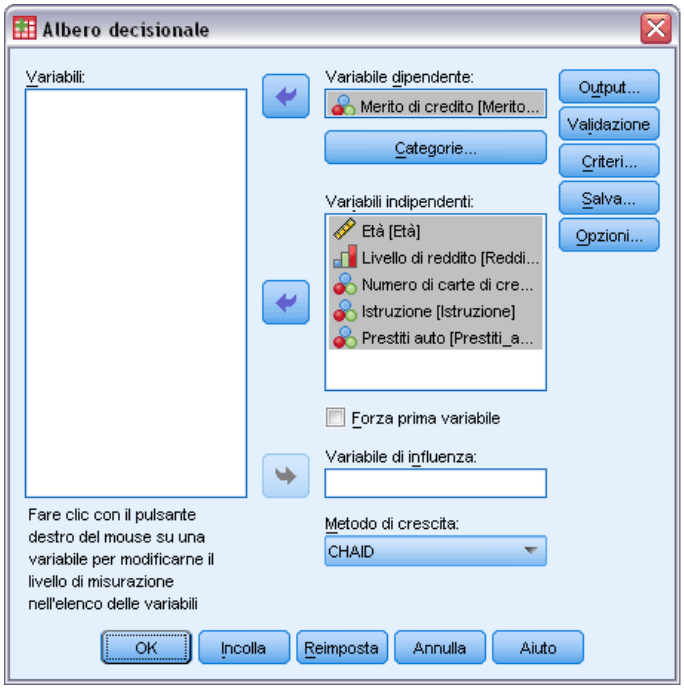
Figura 6-1
Dati di credito con valori mancanti



Come nell'esempio relativo al rischio di credito, (per ulteriori informazioni vedere il capitolo 4), questo esempio illustra il tentativo di creare un modello per classificare i rischi di credito positivi e negativi. La differenza principale è rappresentata dal fatto che questo file di dati include valori mancanti per alcune variabili indipendenti utilizzate nel modello.

- Per eseguire un'analisi Albero decisionale, dai menu scegliere:
  Analizza > Classifica > Albero...

Figura 6-2
Finestra di dialogo Albero decisionale

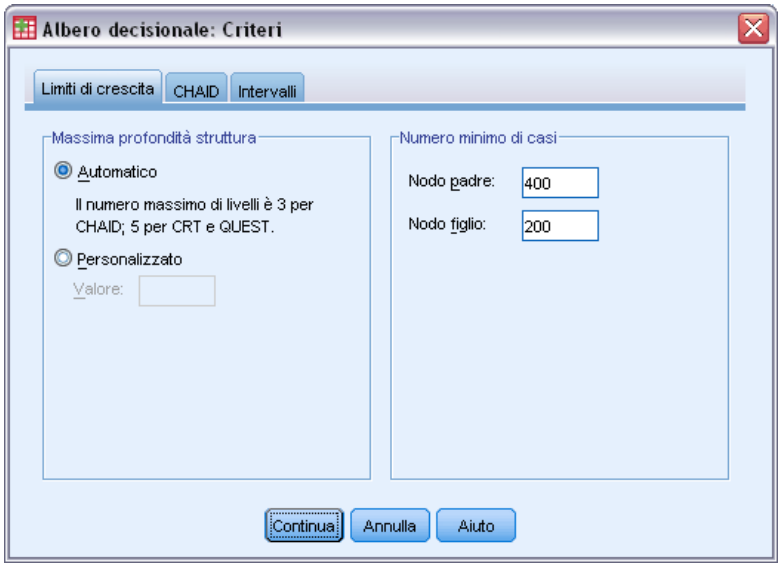


- Selezionare *Valutazione credito* come variabile dipendente.
- Selezionare tutte le variabili rimanenti come indipendenti (la procedura escluderà automaticamente eventuali variabili che non contribuiscano in modo significativo al modello finale).
- Come metodo di espansione scegliere CHAID.

  Nell'esempio corrente si vuole mantenere l'albero piuttosto semplice; per questo si limiterà l'espansione dell'albero aumentando il numero minimo di casi per i nodi genitore e i nodi figlio.

- Nella finestra di dialogo principale Albero decisionale fare clic su Criteri.

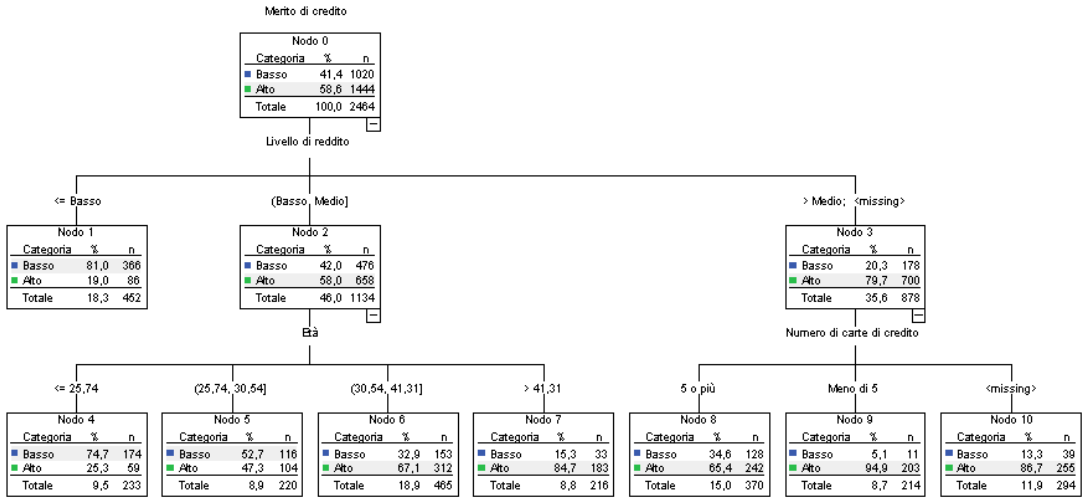
Figura 6-3
Finestra di dialogo Criteri, scheda Limiti di crescita



- In Numero minimo di casi, digitare 400 per Nodo genitore e 200 per Nodo figlio.
- Fare clic su Continua e quindi su OK per eseguire la procedura.

Risultati CHAID

Figura 6-4
Albero CHAID con valori delle variabili indipendenti mancanti



Per il nodo 3, il valore di *livello di reddito* è visualizzato come *>Medio;<mancante>*. Questo significa che il nodo include casi nella categoria a reddito elevato più alcuni casi con valori mancanti per *livello di reddito*.

Il nodo terminale 10 include casi con valori mancanti per *numero di carte di credito*. Per identificare i rischi di credito positivi, questo è in realtà il secondo nodo terminale migliore, il che può essere problematico se si desidera utilizzare il modello corrente per prevedere rischi di credito positivi. È probabile che non si desideri un modello che preveda una valutazione del credito positiva basata sul fatto che non è noto il numero di carte di credito possedute da un caso e sul fatto che per alcuni di questi casi mancano informazioni sul livello di reddito.

Figura 6-5
Tabelle di rischio e di classificazione per il modello CHAID

Rischio

Stima	Errore standard
,249	,009

Metodo di crescita: CHAID
Variabile dipendente: Merito di credito

Classificazione

	Previsione		
Osservato	Basso	Alto	Percentuale corretta
Basso	656	364	64,3%
Alto	249	1195	82,8%
Percentuale globale	36,7%	63,3%	75,1%

Metodo di crescita: CHAID
Variabile dipendente: Merito di credito

Tabelle di rischio e di classificazione indicano che il modello CHAID classifica correttamente circa il 75% dei casi. Si tratta di una percentuale non eccellente. Inoltre, è possibile che siano motivi di sospettare che il tasso di classificazione corretta per i casi di credito positivo possa essere eccessivamente ottimistico, in quanto basato in parte sul presupposto che l'assenza di informazioni su due variabili indipendenti (*livello di reddito* e *numero di carte di credito*) sia indicativa di una posizione creditizia positiva.

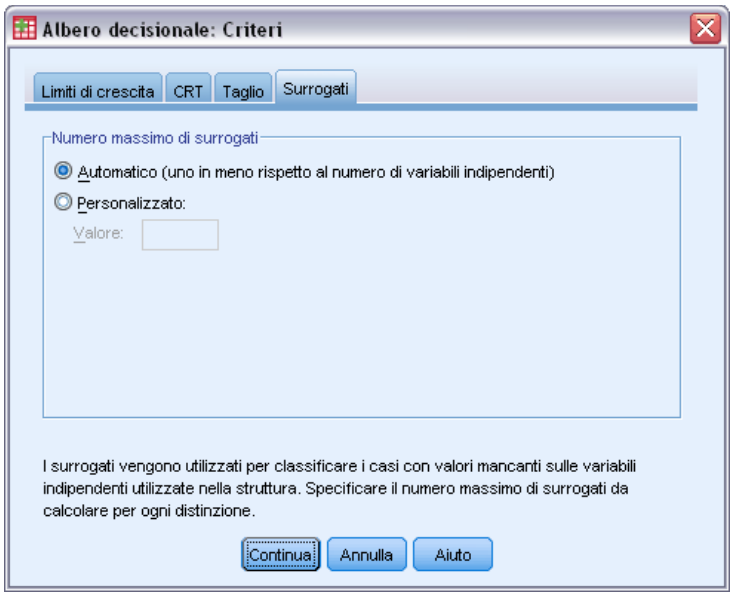
Valori mancanti con CRT

Si proverà ora ad applicare la stessa analisi di base, ma utilizzando CRT come metodo di espansione.

- Nella finestra di dialogo principale Albero decisionale fare clic su CRT come metodo di espansione.
- Fare clic su Criteri.
- Assicurarsi che il numero minimo di casi sia impostato su 400 per i nodi genitore e su 200 i nodi figlio.
- Fare clic sulla scheda Surrogati.

  Nota: La scheda Surrogati sarà disponibile solo se è stato selezionato CRT o QUEST come metodo di espansione.

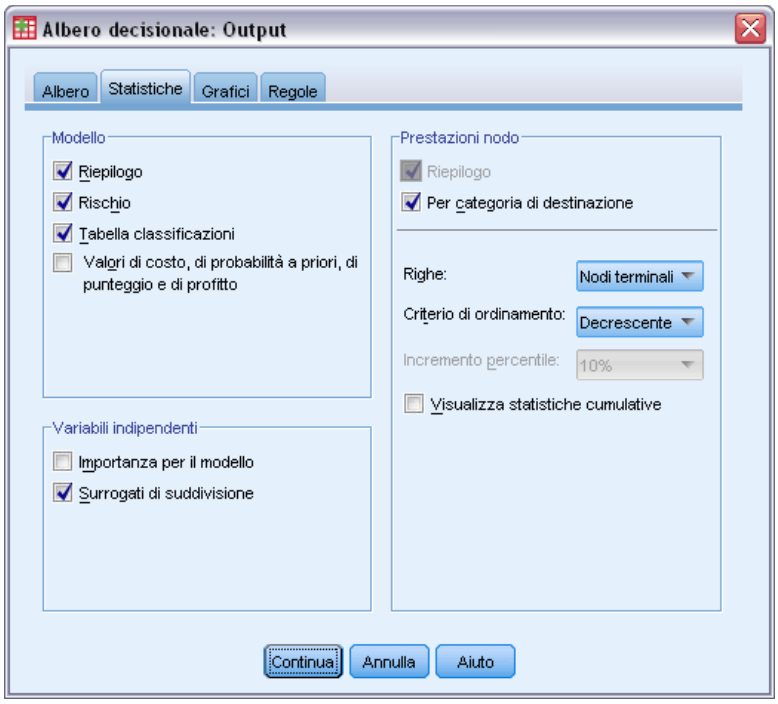
Figura 6-6
Finestra di dialogo Criteri, scheda Surrogati



Per ciascuna divisione del nodo delle variabili indipendenti, l'impostazione Automatico considererà ogni altra variabile indipendente specificata per il modello come possibile surrogato. Poiché nell'esempio corrente il numero delle variabili indipendenti non è elevato, l'impostazione Automatico risulta appropriata.

- Fare clic su Continua.
- Nella finestra di dialogo principale Albero decisionale fare clic su Output.

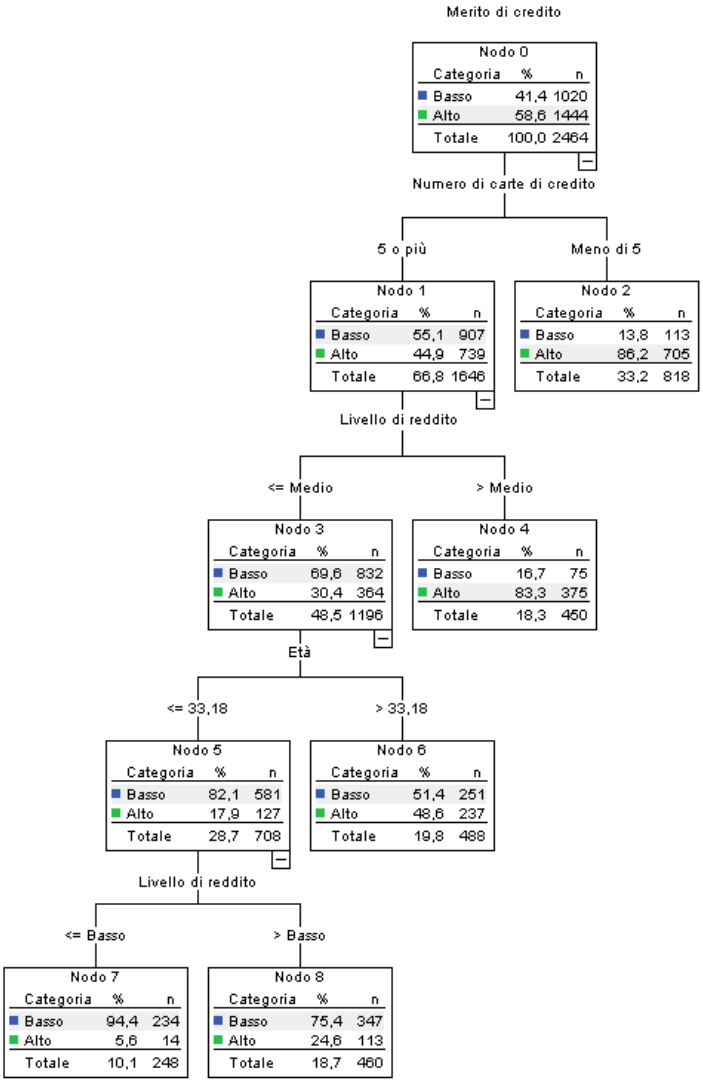
Figura 6-7
Finestra di dialogo Output, scheda Statistiche



- Fare clic sulla scheda Statistiche.
- Selezionare Surrogati di suddivisione.
- Fare clic su Continua e quindi su OK per eseguire la procedura.

Risultati CRT

Figura 6-8
Albero CRT con valori delle variabili indipendenti mancanti



Si noterà immediatamente che l'albero corrente non è molto simile all'albero CHAID. Questo di per sé non è molto significativo. In un modello di albero CRT, tutte le divisioni sono binarie, ovvero ogni nodo genitore viene diviso in due soli nodi figlio. In un modello CHAID, i nodi genitore possono essere divisi in molti nodi figlio. In questo modo, gli alberi avranno spesso aspetti diversi anche se rappresentano lo stesso modello sottostante.

Esistono, tuttavia, diverse importanti differenze:

- La variabile (predittrice) indipendente più importante nel modello CRT è *numero di carte di credito*, mentre nel modello CHAID il predittore più importante era *livello di reddito*.

- Per casi con meno di cinque carte di credito, *numero di carte di credito* è l'unico predittore significativo della valutazione del credito e il nodo 2 è un nodo terminale.
- Come nel modello CHAID, anche *livello di reddito* ed *età* sono inclusi nel modello, sebbene *livello di reddito* sia ora il secondo predittore anziché il primo.
- Non sono presenti nodi che includano una categoria <*mancante*>, in quanto CRT utilizza predittori surrogati anziché valori mancanti nel modello.

Figura 6-9
Tabelle di rischio e di classificazione per il modello CRT

Rischio

Stima	Errore standard
,224	,008

Metodo di crescita: CRT
Variabile dipendente: Merito di credito

Classificazione

	Previsione		
Osservato	Basso	Alto	Percentuale corretta
Basso	832	188	81,6%
Alto	364	1080	74,8%
Percentuale globale	48,5%	51,5%	77,6%

Metodo di crescita: CRT
Variabile dipendente: Merito di credito

- Le tabelle di rischio e di classificazione mostrano un tasso di classificazione corretta globale pari a quasi il 78%, con un lieve aumento rispetto al modello CHAID (75%).
- Il tasso di classificazione corretta per i casi di credito negativo è molto più alta per il modello CRT—81,6% rispetto al 64,3% del modello CHAID.
- Il tasso di classificazione corretta per i casi di credito positivo, tuttavia, è sceso dall'82,8% del modello CHAID al 74,8% del modello CRT.

Surrogati

Le differenze tra i modelli CHAID e CRT sono dovute, in parte, all'utilizzo di surrogati nel modello CRT. La tabella dei surrogati ne indica la modalità di utilizzo nel modello.

Figura 6-10
Tabella Surrogati

Nodo padre	Variabile dipendente		Miglioramento	Associazione
0	Principale	Numero di carte di credito	,090	
	Surrogate	Prestiti auto	,052	,643
		Età	,001	,004
1	Principale	Livello di reddito	,071	
	Surrogate	Età	,001	,004
3	Principale	Età	,022	
5	Principale	Livello di reddito	,006	
	Surrogate	Età	,000	,009

Growing Method: CRT
Dependent Variable: Merito di credito

- A livello di nodo radice (nodo 0), la variabile (predittrice) indipendente migliore è *numero di carte di credito*.
- Per i casi con valori mancanti per *numero di carte di credito*, *finanziamenti auto* è utilizzata come predittore surrogato, poiché la variabile ha un'associazione piuttosto elevata (0,643) con *numero di carte di credito*.
- Se in un caso è presente un valore mancante anche per *finanziamenti auto*, come surrogato viene utilizzata *età* (anche se il suo valore di associazione, 0,004, è piuttosto ridotto).
- *Età* è utilizzata come surrogato anche per *livello di reddito* per i nodi 1 e 5.

Riepilogo

Metodi di espansione diversi gestiscono i dati mancanti in modi diversi. Se i dati utilizzati per creare il modello includono molti valori mancanti—o se si desidera applicare il modello ad altri file di dati che contengono molti valori mancanti—sarà necessario valutare l'effetto dei valori mancanti sui diversi modelli. Se si desidera utilizzare i surrogati nel modello per compensare i valori mancanti, utilizzare i modelli CRT o QUEST.

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 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 , 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 , 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 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 , 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 . 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" . 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 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 ha presentato una tabella per illustrare i possibili gruppi sociali. Guttman 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 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, e , 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 .
- **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 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 .
- **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 .
- **verd1985.sav.** Questo file di dati prende in esame un'indagine . 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 . 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.

Indice

alberi, 1
  albero in formato tabella, 67
  applicazione di modelli, 83
  attributi del testo, 44
  caratteri, 44
  colori, 44
  colori dei grafici dei nodi, 44
  contenuto dell'albero in una tabella, 25
  controllo della visualizzazione dell'albero, 25, 43
  controllo delle dimensioni dei nodi, 9
  convalida con suddivisione, 8
  convalida incrociata, 8
  costi di errata classificazione, 17
  costi personalizzati, 78
  Criteri di espansione CHAID, 10
  effetti del livello di misurazione, 50
  effetti delle etichette dei valori, 54
  generazione di regole, 37, 46
  grafici, 31
  guadagni per la tabella nodi, 69
  importanza predittore, 27
  intervalli per variabili indipendenti di scala, 11
  limitazione del numero di livelli, 9
  mancata visualizzazione di rami e nodi, 39
  mappa dell'albero, 41
  metodo CRT, 12
  modifica, 39
  mostrare e nascondere le statistiche relative ai rami, 25
  orientamento dell'albero, 25
  probabilità a priori, 19
  profitti, 18
  punteggi, 21
  punteggio, 83
  salvataggio di valori attesi, 73
  salvataggio di variabili di modello, 24
  scaling della visualizzazione dell'albero, 42
  selezione di casi nei nodi, 74
  selezione di più nodi, 39
  statistiche dei nodi terminali, 27
  stime del rischio, 27
  stime del rischio per variabili dipendenti di scala, 88
  surrogati, 93, 100
  tabella di errata classificazione, 27
  tabella Riepilogo del modello, 65
  taglio, 15
  utilizzo di alberi di grandi dimensioni, 40
  valori indice, 27
  valori mancanti, 22, 93
  variabili dipendenti di scala, 83
alberi decisionali , 1
  forzatura della prima variabile nel modello, 1
  livello di misurazione, 1
  metodo CHAID, 1
  metodo CHAID esaustivo, 1
  metodo CRT, 1
  metodo QUEST, 1, 14

CHAID, 1
  Correzione di Bonferroni, 10
  criteri di unione e di divisione, 10
  intervalli per variabili indipendenti di scala, 11
  massimo numero di iterazioni, 10
  ridivisione di categorie unite, 10
compressione dei rami dell'albero, 39
convalida
  alberi, 8
convalida con suddivisione
  alberi, 8
convalida incrociata
  alberi, 8
costi
  errata classificazione, 17
  modelli di albero, 78
CRT, 1
  misure di impurità, 12
  taglio, 15

errata classificazione
  alberi, 27
  costi, 17
  tassi, 72
etichette dei valori
  alberi, 54

file di esempio
  posizione, 103

Gini, 12
grafico degli indici, 71
grafico guadagni, 70
guadagno, 69

impurità
  Alberi CRT, 12
Indice
  modelli di albero, 69

legal notices, 114
livello di misurazione
  alberi decisionali, 1
  nei modelli di albero, 50
livello di significatività per la divisione dei nodi, 14

mancata visualizzazione di rami dell'albero, 39
modelli di albero, 69

nascondere i nodi
  confronto con taglio, 15
nodi
  selezione di più nodi dell'albero, 39
numero dei nodi
  salvataggio come variabile da alberi decisionali, 24

peso di casi
  pesi frazionari negli alberi decisionali, 1
probabilità prevista
  salvataggio come variabile da alberi decisionali, 24
profitti
  alberi, 18, 27
  probabilità a priori, 19
punteggi
  alberi, 21
punteggio
  modelli di albero, 83

QUEST, 1, 14
  taglio, 15

regole
  creazione di sintassi di selezione e punteggio per alberi decisionali, 37, 46
risposta
  modelli di albero, 69

selezione di più nodi dell'albero, 39
seme dei numeri casuali
  convalida alberi decisionali, 8
sintassi
  creazione di sintassi di selezione e punteggio per alberi decisionali, 37, 46
sintassi dei comandi
  creazione di sintassi di selezione e punteggio per alberi decisionali, 37, 46
SQL
  creazione di sintassi SQL per selezione e punteggio, 37, 46
stime del rischio
  alberi, 27
  per variabili dipendenti categoriali, 72
  per variabili dipendenti di scala nella procedura Albero decisionale, 88
surrogati
  nei modelli di albero, 93, 100

tabella di classificazione, 72
tabella Riepilogo del modello
  modelli di albero, 65
taglio alberi decisionali
  confronto con nascondere i nodi, 15
trademarks, 115
twoing, 12
twoing ordinato, 12

valori attesi
  salvataggio come variabile da alberi decisionali, 24
  salvataggio per modelli di albero, 73
valori indice
  alberi, 27
valori mancanti
  alberi, 22
  nei modelli di albero, 93
variabili di scala
  variabili dipendenti nella procedura Albero decisionale, 83