

*IBM SPSS Decision Trees 29*



**Nota**

Prima di utilizzare queste informazioni e il prodotto che supportano, leggere le informazioni in [“Informazioni particolari” a pagina 21](#).

**Informazioni sul prodotto**

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

© **Copyright International Business Machines Corporation .**

---

# Indice

<b>Capitolo 1. Alberi decisionali.....</b>	<b>1</b>
Creazione Alberi di decisione.....	1
Selezione Categorie.....	4
Convalida.....	4
Albero - Criteri di coltivazione.....	5
Opzioni.....	8
Salvataggio delle informazioni sul modello.....	11
Output.....	12
Editor degli alberi.....	16
Lavorare con Large Trees.....	17
Controllo Informazioni Visualizzate nell'Albero.....	18
Modifica Colori dell'albero e font di testo.....	18
Regole di selezione e di valutazione dei casi.....	19
<b>Informazioni particolari.....</b>	<b>21</b>
Marchi.....	22
<b>Indice analitico.....</b>	<b>25</b>



---

# Capitolo 1. Alberi decisionali

Le seguenti funzioni di decisione Trees sono incluse in SPSS Statistiche Professional Edition o l'opzione Decisione Trees.

---

## Creazione Alberi di decisione

La procedura dell'albero di decisione crea un modello di classificazione basato sull'albero. Classifica i casi in gruppi o prevede valori di una variabile dipendente (target) basata su valori di variabili indipendenti (predictor). La procedura fornisce strumenti di validazione per analisi di classificazione esplorativa e di conferma.

La procedura può essere utilizzata per:

**Segmentazione.** Identificare le persone suscettibili di essere membri di un determinato gruppo.

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

**Previsione.** Creare regole e usarle per prevedere gli eventi futuri, come la probabilità che qualcuno di default su un prestito o il valore potenziale di rivendita di un veicolo o di una casa.

**riduzione dei dati e screening variabile.** Seleziona un sottoinsieme utile di predittori per un insieme ampio di variabili da utilizzare nella creazione di un modello parametrico formale.

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

**Categoria di unione e di variabili continue discretizing.** Ricodificare le categorie di predittori di gruppo e le variabili continue con una perdita minima di informazioni.

**Esempio.** Una banca vuole catalogare i richiedenti del credito a seconda che rappresentino o meno un rischio di credito ragionevole. In base a vari fattori, tra cui i noti rating di credito dei clienti passati, è possibile costruire un modello per prevedere se i futuri clienti saranno probabilmente inadempienti sui loro prestiti.

Un'analisi basata sull'albero fornisce alcune funzioni attraenti:

- Permette di identificare gruppi omogenei ad alto o basso rischio.
- Rende facile costruire regole per fare previsioni sui singoli casi.

Considerazioni sui dati

**Dati.** Le variabili dipendenti e indipendenti possono essere:

- *Nominale.* Una variabile può essere considerata nominale quando i suoi valori rappresentano categorie prive di classificazione intrinseca (ad esempio, il dipartimento della società in cui lavora un dipendente). Degli esempi di variabili nominali includono la regione, il codice postale e l'affiliazione religiosa.
- *Ordinale.* Una variabile può essere trattata come ordinale quando i suoi valori rappresentano categorie con una classificazione intrinseca (ad esempio, livelli di soddisfazione del servizio da molto insoddisfatti a molto soddisfatti). Degli esempi di variabili ordinali includono i punteggi di atteggiamento che rappresentano i gradi di soddisfazione o fiducia e i punteggi di classificazione delle preferenze.
- *Scala.* Una variabile può essere trattata come una scala (continua) quando i suoi valori rappresentano categorie ordinate con una metrica significativa, in modo che i confronti di distanza tra i valori siano appropriati. Esempi di variabili di scala sono l'età espressa in anni o il reddito espresso in migliaia di Euro.

**Pesi di frequenza** Se la ponderazione è in vigore, i pesi frazionali sono arrotondati al numero intero più vicino; quindi, i casi con un valore di peso inferiore a 0,5 vengono assegnati un peso di 0 e sono quindi esclusi dall'analisi.

**Ipotesi.** Questa procedura presuppone che il livello di misurazione appropriato sia stato assegnato a tutte le variabili di analisi e alcune funzioni ipotizza che tutti i valori della variabile dipendente inclusi nell'analisi abbiano definito etichette di valore.

- **Livello di misurazione.** Il livello di misurazione influisce sulle elaborazioni degli alberi; quindi, a tutte le variabili va assegnato il livello di misurazione appropriato. Per impostazione predefinita, si ipotizza che le variabili numeriche siano variabili e le variabili di stringa sono considerate nominali, che possono non riflettere accuratamente il vero livello di misurazione. Un'icona accanto ad ogni variabile nell'elenco delle variabili identifica il tipo di variabile.

Tabella 1. Icone del livello di misurazione

Icona	Livello di misurazione
	Scala
	Nominale
	Ordinale

È possibile modificare temporaneamente il livello di misurazione per una variabile facendo clic con il tasto destro del mouse sulla variabile nell'elenco delle variabili di origine e selezionando un livello di misurazione dal menu a comparsa.

- **Etichette valori.** L'interfaccia della finestra di dialogo per questa procedura presuppone che tutti i valori non mancanti di una variabile dipendente categoriale (nominale, ordinale) abbiano definito etichette di valore o nessuna di esse. Alcune funzioni non sono disponibili a meno che almeno due valori non mancanti della variabile dipendente categoriale abbiano etichette di valore. Se almeno due valori non mancanti hanno definito etichette di valore, eventuali casi con altri valori che non hanno etichette di valore saranno esclusi dall'analisi.

Per Ottenere Alberi Decisionali

1. Dai menu, scegliere:

**Analizzare > Classifica > Albero ...**

2. Selezionare una variabile dipendente.
3. Selezionare una o più variabili indipendenti.
4. Selezionare un metodo di coltivazione.

Facoltativamente, è possibile:

- Modificare il livello di misurazione per qualsiasi variabile nella lista sorgente.
- Forza la prima variabile nell'elenco delle variabili indipendenti nel modello come prima variabile split.
- Selezionare una variabile di influenza che definisce quanto influenza un caso sul processo di coltivazione dell'albero. I casi con valori di influenza più bassi hanno meno influenza; i casi con valori più alti ne hanno di più. Influenzare i valori variabili deve essere positivo.
- Convalidare l'albero.
- Personalizzare i criteri di crescita dell'albero.
- Salvare i numeri del nodo terminale, i valori previsti 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) del dataset è sconosciuto. Poiché influisce sul calcolo dei risultati di questa procedura, il livello di misurazione deve essere definito per tutte le variabili.

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

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

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

Modifica del livello di misurazione

1. Fare clic con il tasto destro del mouse sulla variabile nell'elenco di origine
2. Selezionare un livello di misurazione dal menu a comparsa.

Ciò modifica il livello di misurazione temporaneamente per l'utilizzo nella procedura dell'Albero della decisione.

Metodi Crescenti

I metodi di coltivazione disponibili sono:

**CHAID.** Rilevamento interazione automatico chi - quadrato. Per ogni fase, 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.** Strutture ad albero di classificazione e regressione. CRT suddivide 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 "puro".

**QUEST.** QUEST (Quick, Unbiased, Efficient Statistical Tree). Metodo che esegue i calcoli molto velocemente ed evita la distorsione degli altri metodi a favore dei predittori con molte categorie QUEST può essere specificato solo se la variabile dipendente è nominale.

Ci sono vantaggi e limitazioni con ogni metodo, tra cui:

Funzione	CHAID*	CRT	QUEST
Chi - quadrato - based**	X		
Variabili indipendenti (predictor) surrogate		X	X
Riduzione della struttura ad albero		X	X
Suddivisione del nodo multiway	X		
Suddivisione del nodo binario		X	X
Variabili di influenza	X	X	
Probabilità a priori		X	X
Costi di errata classificazione	X	X	X
Calcolo rapido	X		X

\*Include Esaurienti CHAID.

\*\* QUEST utilizza anche una misura chi quadrato per le variabili indipendenti nominali.

## Selezione Categorie

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

- Controllo quali categorie sono incluse nell'analisi.
- Identificare le categorie target di interesse.

Incluse / Escluse Categorie

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

- I casi con valori della variabile dipendente nell'elenco Escludi non sono inclusi nell'analisi.
- Per le variabili dipendenti nominali è possibile includere anche le categorie mancanti dell'utente nell'analisi. (Per impostazione predefinita, le categorie mancanti dell'utente sono visualizzate nell'elenco Escludi).

Categorie Target

Le categorie selezionate (controllate) sono trattate come le categorie di interesse primario nell'analisi. Ad esempio, se si è interessati principalmente a identificare quegli individui più inclini a default su un prestito, si potrebbe selezionare la categoria "bad" credit - rating come categoria di destinazione.

- Non esiste una categoria di destinazione predefinita. Se non viene selezionata alcuna categoria, non sono disponibili alcune opzioni di regola di classificazione e l'output relativo ai guadagni.
- Se vengono selezionate più categorie, vengono prodotte tabelle e grafici di guadagni separati per ogni categoria di destinazione.
- Designare una o più categorie come categorie target non ha alcun effetto sul modello di albero, sulla stima del rischio o sui risultati di misclassificazione.

Categorie e Value Labels

Questa finestra di dialogo richiede etichette di valore definite per la variabile dipendente. Non è disponibile a meno che almeno due valori della variabile dipendente categoriale abbiano definito etichette di valore.

Per Includere / Escludere Categorie e Seleziona categorie di destinazione

1. Nella finestra di dialogo Albero principale della decisione selezionare una variabile dipendente categoriale (nominale, ordinale) con due o più etichette di valore definite.
2. Fare clic su **Categorie**.

## Convalida

La validazione consente di valutare quanto la struttura dell'albero generalizzi a una popolazione più ampia. Sono disponibili due metodi di validazione: validazione crossvalidation e split - sample.

Convalida incrociata

Crossvalidation divide il campione in un numero di sottocampioni, o **pieghe**. I modelli di albero vengono poi generati, escludendo i dati di ciascun sottocampione a turno. Il primo albero si basa su tutti i casi tranne quelli nella prima piega del campione, il secondo albero si basa su tutti i casi tranne quelli nella seconda piega campione, e così via. Per ogni albero il rischio di misclassificazione è stimato applicando l'albero al sottocampione escluso nel generarlo.

- È possibile specificare un massimo di 25 pieghe di esempio. Più alto è il valore, minore è il numero di casi esclusi per ogni modello di albero.
- Crossvalidation produce un unico modello di albero finale. La stima del rischio incrociato per l'albero finale è calcolata come la media dei rischi per tutti gli alberi.

Convalida del campione suddiviso

Con la validazione del campione split, il modello viene generato utilizzando un campione di allenamento e testato su un campione di hold-out.

- È possibile specificare una dimensione del campione di formazione, espressa in percentuale rispetto alla dimensione totale del campione, oppure una variabile che suddivide il campione in campioni di formazione e verifica.
- Se si utilizza una variabile per definire i campioni di formazione e verifica, i casi con un valore di 1 per la variabile vengono assegnati al campione di formazione e tutti gli altri casi sono assegnati al campione di prova. 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 sia per la formazione che per i campioni di prova o solo per il campione di prova.
- La convalida del campione split deve essere utilizzata con cautela su piccoli file di dati (file di dati con un numero ridotto di casi). Le piccole dimensioni del campione di allenamento possono produrre modelli poveri, dato che potrebbero non esserci abbastanza casi in alcune categorie per far crescere adeguatamente l'albero.

Per convalidare un albero di decisione

1. Nella finestra di dialogo Trees principale, clicca su **Validation**.
2. Selezionare **Crossvalidation** o **Convalidazione a campione di Split**.

*Nota:* Entrambi i metodi di validazione assegnano casualmente i casi ai gruppi di esempio. Se si desidera essere in grado di riprodurre gli stessi identici risultati in una successiva analisi, è necessario impostare il seme del numero casuale (Menu Trasforma, Random Number Generators) prima di eseguire l'analisi per la prima volta e poi resettare il seme a tale valore per la successiva analisi.

## Albero - Criteri di coltivazione

I criteri di coltivazione disponibili possono dipendere dal metodo di coltivazione, dal livello di misurazione della variabile dipendente, o da una combinazione dei due.

### Limiti di crescita:

La Scheda Limiti di crescita consente di limitare il numero di livelli nell'albero e di controllare il numero minimo di casi per i nodi padre e figlio.

**Massima profondità dell'albero.** Controlla il numero massimo di livelli di crescita sotto il nodo root. L'impostazione **Automatica** limita l'albero a tre livelli sotto il nodo root per i metodi CHAID e CHAID esaustivi e cinque livelli per i metodi CRT e QUEST.

**Numero minimo di casi.** Controlla i numeri minimi dei casi per i nodi. I nodi che non soddisfano questi criteri non saranno suddivisi.

- Aumentare i valori minimi tende a produrre alberi con meno nodi.
- Diminuire i valori minimi produce alberi con più nodi.

Per i file di dati con un numero ridotto di casi, i valori predefiniti di 100 casi per i nodi principali e di 50 casi per i nodi child possono talvolta comportare alberi senza nodi al di sotto del nodo root; in questo caso, abbassando i valori minimi possono produrre risultati più utili.

Per Specificare Limiti di crescita

1. Nella finestra di dialogo Albero della decisione principale, fare clic su **Criteri**.
2. Fare clic sulla scheda **Limiti di crescita**.

### Criteri CHAID

Per i metodi CHAID e CHAID esaustivi è possibile controllare:

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

- Per la suddivisione dei nodi il valore deve essere superiore a 0 e inferiore a 1. Valori minori tendono a generare strutture ad albero con un numero inferiore di nodi.
- Per le categorie di fusione il valore deve essere superiore a 0 e inferiore o uguale a 1. Per impedire la fusione di categorie, specificare un valore di 1. Per una variabile indipendente in scala, ciò significa che il numero di categorie per la variabile nell'albero finale è il numero di intervalli specificato (il default è 10). Per ulteriori informazioni, consultare la sezione “Intervalli Di Scala per l'analisi CHAID” a pagina 6.

**Chi - Square Statistica.** Per le variabili dipendenti ordinali, chi - quadrato per determinare la suddivisione del nodo e la fusione di categoria viene calcolato utilizzando il metodo di probabilità - rapporto. Per le variabili dipendenti nominali è possibile selezionare il metodo:

- **Pearson.** Questo metodo fornisce calcoli più rapidi ma deve essere utilizzato con cautela su campioni di piccole dimensioni. 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'albero smette di crescere perché è stato raggiunto il numero massimo di iterazioni, si può voler aumentare il massimo o modificare uno o più degli altri criteri che controllano la crescita dell'albero.
- **Modifica minima nelle frequenze di cella previste.** Il valore deve essere superiore a 0 e inferiore a 1. Il valore predefinito è 0.05. Valori minori tendono a generare strutture ad albero con un numero inferiore di nodi.

**Adegua valori di significatività tramite il metodo di Bonferroni.** Per confronti multipli, i valori di significatività per la fusione e la suddivisione dei criteri vengono regolati utilizzando il metodo Bonferroni. È l'impostazione predefinita.

**Consenti la risuddivisione delle categorie unite all'interno di un nodo.** A meno che non si impedisca esplicitamente la fusione di categoria, la procedura tenterà di unire le categorie variabili indipendenti (predittore) per produrre l'albero più semplice che descrive il modello. Questa opzione consente la procedura alle categorie fuse unite se questo fornisce una soluzione migliore.

A specificare i Criteri CHAID

1. Nella finestra di dialogo Albero della decisione principale, selezionare **CHAID** o **CHAID esaustivo** come metodo di coltivazione.
2. Fare clic su **Criteri**.
3. Fare clic sulla scheda **CHAID** .

### ***Intervalli Di Scala per l'analisi CHAID***

In CHAID analisi, le variabili indipendenti (predittore) sono sempre bandate in gruppi discreti (ad esempio, 0-10, 11-20, 21-30, ecc.) prima dell'analisi. È possibile controllare il numero iniziale / massimo di gruppi (sebbene la procedura possa unire gruppi contigui dopo la suddivisione iniziale):

- **Numero fisso.** Tutte le variabili indipendenti di scala sono inizialmente bandate nello stesso numero di gruppi. Il valore predefinito è 10.
- **Personalizzato.** Ogni variabile indipendente di scala viene inizialmente bandito nel numero di gruppi specificati per quella variabile.

Per Specificare Intervalli per Variabili Indipendenti

1. Nella finestra di dialogo Albero della decisione principale, selezionare una o più variabili indipendenti di scala.
2. Per il metodo di coltivazione, selezionare **CHAID** o **CHAID esaustivo**.
3. Fare clic su **Criteri**.
4. Fare clic sulla scheda **Intervals** .

Nelle analisi CRT e QUEST, tutte le divise sono binarie e scale e le variabili indipendenti ordinarie vengono gestite allo stesso modo; quindi, non è possibile specificare un numero di intervalli per variabili indipendenti di scala.

## Criteri CRT

Il metodo di coltivazione 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'indicazione di **impurezza**. 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 ulteriore suddivisione perché è "puro".

È possibile selezionare il metodo utilizzato per misurare l'impurità e la diminuzione minima di impurità necessaria per suddividere i nodi.

**Misura di impurità.** Per le variabili dipendenti in scala si utilizza la misura di deviazione meno squadrata (LSD) di impurità. Viene calcolato come varianza all'interno del nodo, regolata per eventuali pesi di frequenza o valori di influenza.

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

- **Gini.** Si trovano suddivise che massimizzano l'omogeneità dei nodi child rispetto al valore della variabile dipendente. Gini si basa su probabili probabilità di appartenenza per ogni categoria della variabile dipendente. Raggiunge il suo minimo (zero) quando tutti i casi in un nodo rientrano in una sola categoria. Questa è la misura predefinita.
- **Twoing.** Le categorie della variabile dipendente sono raggruppate in due sottoclassi. È stato rilevato che le suddivisioni separano meglio i due gruppi.
- **Ordinato twoing.** Simile al twoing tranne che solo le categorie adiacenti possono essere raggruppate. Questa misura è disponibile solo per variabili dipendenti ordinali.

**Modifica minima in miglioramento.** Si tratta della diminuzione minima di impurità necessaria per dividere un nodo. Il default è 0,0001. Valori più alti tendono a produrre alberi con meno nodi.

Per Specificare Criteri CRT

1. Per il metodo di coltivazione, selezionare **CRT**.
2. Fare clic su **Criteri**.
3. Fare clic sulla scheda **CRT**.

## Criteri QUEST

Per il metodo QUEST è possibile specificare il livello di significatività per la suddivisione dei nodi. Una variabile indipendente non può essere utilizzata per dividere i nodi a meno che il livello di significatività non sia inferiore o uguale al valore specificato. Il valore deve essere superiore a 0 e inferiore a 1. Il valore predefinito è 0.05. I valori più piccoli tenderanno ad escludere variabili più indipendenti dal modello finale.

Per specificare Criteri QUEST

1. Nella finestra di dialogo Albero della decisione principale, selezionare una variabile dipendente nominale.
2. Per il metodo di coltivazione, selezionare **QUEST**.
3. Fare clic su **Criteri**.
4. Fare clic sulla scheda **QUEST**.

## Alberi Di Pruning

Con i metodi CRT e QUEST è possibile evitare di sovrastare il modello tramite **pruning** l'albero: l'albero è coltivato fino a quando non si arrestano i criteri e poi viene rifilato automaticamente al subalbero più piccolo in base alla differenza massima specificata in rischio. Il valore di rischio è espresso in errori

standard. Il valore predefinito è 1. Il valore deve essere non negativo. Per ottenere il subalbero con il rischio minimo, specificare 0.

Per Prune un Albero

1. Nella finestra di dialogo Albero principale della decisione, per il metodo di coltivazione, selezionare **CRT o QUEST**.
2. Fare clic su **Criteri**.
3. Fare clic sulla scheda **Pruning**.

Pruning versus Nascondendo Nodi

Quando si crea un albero di prato, eventuali snodi eliminati dall'albero non sono disponibili nell'albero finale. Si può interattivamente nascondere e mostrare i nodi child selezionati nell'albero finale, ma non è possibile mostrare nodi inclini al processo di creazione dell'albero. Per ulteriori informazioni, consultare la sezione [“Editor degli alberi” a pagina 16](#).

## Surrogati

CRT e QUEST possono utilizzare **surrogati** per variabili indipendenti (predittore). Per i casi in cui manca il valore per tale variabile, per la classificazione vengono utilizzate altre variabili indipendenti che presentano alte associazioni con la variabile originale. Questi predittori alternativi sono chiamati surrogati. È possibile specificare il numero massimo di surrogati da utilizzare nel modello.

- Per impostazione predefinita, il numero massimo di surrogati è uno inferiore al numero di variabili indipendenti. In altre parole, per ogni variabile indipendente, tutte le altre variabili indipendenti possono essere utilizzate come surrogati.
- Se non si desidera che il modello utilizzi surrogati, specificare 0 per il numero di surrogati.

Per Specificare Surrogati

1. Nella finestra di dialogo Albero principale della decisione, per il metodo di coltivazione, selezionare **CRT o QUEST**.
2. Fare clic su **Criteri**.
3. Fare clic sulla scheda **Surrogate**.

## Opzioni

Le opzioni disponibili possono dipendere dal metodo di coltivazione, dal livello di misurazione della variabile dipendente e / o dall'esistenza di etichette di valore definite per valori della variabile dipendente.

## Costi di errata classificazione

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

- Il costo di negare il credito a un cliente degno di credito rischia di essere diverso dal costo di estendere il credito a un cliente che poi inadempie sul prestito.
- Il costo di misclassifying un individuo ad alto rischio di malattie cardiache come basso rischio è probabilmente molto più alto rispetto al costo di misclassificare un individuo a basso rischio come ad alto rischio.
- Il costo di inviare una mailing di massa a qualcuno che probabilmente non risponde è probabilmente abbastanza basso, mentre il costo di non mandare il mailing a qualcuno che probabilmente risponde è relativamente più alto (in termini di entrate perse).

Misclassifica Costi e Value Labels

Questa finestra di dialogo non è disponibile a meno che almeno due valori della variabile dipendente categoriale abbiano definito etichette di valore.

Per Specificare I Costi Di Errata Classificazione

1. Nella finestra di dialogo Albero principale della decisione selezionare una variabile dipendente categoriale (nominale, ordinale) con due o più etichette di valore definite.
2. Fare clic su **Opzioni**.
3. Fare clic sulla scheda **Costi di misclassificazione** .
4. Fare clic su **Personalizza**.
5. Inserire uno o più costi di errata classificazione nella griglia. I valori devono essere non negativi. (Le classificazioni corrette, rappresentate sulla diagonale, sono sempre 0.)

**Riempita matrice.** In molte istanze, si potrebbe desiderare dei costi per essere simmetrici - vale a dire il costo di misclassifying A come B è lo stesso del costo di misclassifying B come A. I seguenti controlli possono facilitare la specifica di una matrice di costo simmetrica:

- **Triangolo inferiore duplicato.** Copia i valori nel triangolo inferiore della matrice (sotto la diagonale) nelle corrispondenti celle triangolari superiori.
- **Triangolo superiore duplicato.** Copia i valori nel triangolo superiore della matrice (sopra la diagonale) nelle corrispondenti celle triangolari inferiori.
- **Utilizzare Valori Cellulari Medi.** Per ogni cellula in ogni metà della matrice i due valori (superiore - e inferiore - triangolare) sono mediamente e la media sostituisce entrambi i valori. Ad esempio, se il costo di misclassificazione A come B è 1 e il costo di misclassifying B come A è 3, allora questo controllo sostituisce entrambi i valori con la media  $(1 + 3) / 2 = 2$ .

## Profitti

Per le variabili dipendenti categoriali è possibile assegnare i valori delle entrate e delle spese ai livelli della variabile dipendente.

- Il profitto è calcolato come entrate meno 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 della struttura ad albero di base.
- I valori delle entrate e delle spese devono essere numerici e devono essere specificati per tutte le categorie della variabile dipendente visualizzate nella griglia.

### Profitti e Value Labels

Questa finestra di dialogo richiede etichette di valore definite per la variabile dipendente. Non è disponibile a meno che almeno due valori della variabile dipendente categoriale abbiano definito etichette di valore.

### Per Specificare Profitti

1. Nella finestra di dialogo Albero principale della decisione selezionare una variabile dipendente categoriale (nominale, ordinale) con due o più etichette di valore definite.
2. Fare clic su **Opzioni**.
3. Fare clic sulla scheda **Prof.** .
4. Fare clic su **Personalizza**.
5. Inserire i valori delle entrate e delle spese per tutte le categorie variabili dipendenti elencate nella griglia.

## Probabilità a priori

Per gli alberi CRT e QUEST con variabili dipendenti categoriali è possibile specificare le probabilità precedenti di appartenenza di gruppo. **Le probabilità precedenti** sono stime della frequenza relativa complessiva per ciascuna categoria della variabile dipendente prima di sapere nulla sui valori delle variabili indipendenti (predittore). L'utilizzo delle probabilità precedenti aiuta a correggere qualsiasi crescita dell'albero causata dai dati nel campione che non è rappresentativo dell'intera popolazione.

**Ottenere dal campione di formazione (priori empirici).** Utilizzare questa impostazione se la distribuzione di valori variabili dipendenti nel file dati è rappresentativa della distribuzione della popolazione. Se si utilizza la validazione del campione suddiviso, viene utilizzata la distribuzione dei casi nel campione di formazione.

*Nota:* dato che i casi sono assegnati casualmente al campione di formazione in convalida split - sample, non si conoscerà in anticipo l'effettiva distribuzione dei casi nel campione di formazione. Per ulteriori informazioni, consultare la sezione [“Convalida” a pagina 4.](#)

**Parenti tra le categorie.** Utilizzare questa impostazione se le categorie della variabile dipendente sono rappresentate ugualmente nella popolazione. Ad esempio, se ci sono quattro categorie, circa il 25% dei casi è in ogni categoria.

**Personalizzato.** Inserire un valore non negativo per ogni categoria della variabile dipendente elencata nella griglia. I valori possono essere proporzioni, percentuali, conteggi di frequenza o qualsiasi altro valore che rappresenti la distribuzione dei valori in tutte le categorie.

**Adegua le probabilità a priori utilizzando i costi di errata classificazione.** Se si definiscono costi di misclassificazione personalizzati, è possibile regolare le probabilità precedenti in base a tali costi. Per ulteriori informazioni, consultare la sezione [“Costi di errata classificazione” a pagina 8.](#)

#### Profitti e Value Labels

Questa finestra di dialogo richiede etichette di valore definite per la variabile dipendente. Non è disponibile a meno che almeno due valori della variabile dipendente categoriale abbiano definito etichette di valore.

A specificare le Probabilità di Prior

1. Nella finestra di dialogo Albero principale della decisione selezionare una variabile dipendente categoriale (nominale, ordinale) con due o più etichette di valore definite.
2. Per il metodo di coltivazione, selezionare **CRT** o **QUEST**.
3. Fare clic su **Opzioni**.
4. Fare clic sulla scheda **Probabilità di precedenza**.

## 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 diminuire la distanza relativa tra valori ordinali o per modificare l'ordine dei valori.

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

Esempio

<i>Tabella 3. Valori di punteggio personalizzati</i>		
<b>Etichetta valore</b>	<b>Valore originale</b>	<b>Punteggio</b>
Non qualificati	1	1
Manuale qualificato	2	4
Impiegato	3	4.5
Professionale	4	7
Gestione	5	6

- I punteggi aumentano la distanza relativa tra *Unqualificati* e *Succiso manuale* e diminuiscono la distanza relativa tra *manuale Succiso* e *Clerical*.
- I punteggi inverte l'ordine di *Gestione* e *Professional*.

#### Punteggi e Value Labels

Questa finestra di dialogo richiede etichette di valore definite per la variabile dipendente. Non è disponibile a meno che almeno due valori della variabile dipendente categoriale abbiano definito etichette di valore.

#### Per Specificare Punteggi

1. Nella finestra di dialogo Albero principale della decisione selezionare una variabile dipendente ordinale con due o più etichette di valore definite.
2. Per il metodo di coltivazione, selezionare **CHAID** o **CHAID esaustivo**.
3. Fare clic su **Opzioni**.
4. Fare clic sulla scheda **Scorci**.

### Valori mancanti

La scheda Valori mancanti controlla la gestione dei valori variabili nominali, mancanti di utenti, indipendenti (predittori).

- La gestione di valori variabili indipendenti dall'utente ordinale e in scala varia tra i metodi di coltivazione.
- La gestione delle variabili dipendenti nominali viene specificata nella finestra di dialogo Categorie. Per ulteriori informazioni, consultare la sezione [“Selezione Categorie”](#) a pagina 4.
- Per variabili dipendenti ordinali e di scala, sono sempre esclusi i casi con valori variabili dipendenti a scomparsa di sistema o all'utente.

**Trattare come valori mancanti.** I valori mancanti dell'utente sono trattati come valori mancanti di sistema. La gestione dei valori mancanti di sistema varia tra i metodi di coltivazione.

**Trattare come valori validi.** I valori mancanti dell'utente delle variabili indipendenti nominali sono trattati come valori ordinari nell'albero in crescita e nella classificazione.

#### Metodo - Norme dipendenti

Se alcuni, ma non tutti, i valori variabili indipendenti sono di sistema - o di user - missing:

- Per CHAID e Exhaustive CHAID, i valori variabili indipendenti di sistema e quelli mancanti di utente sono inclusi nell'analisi come una categoria unica e combinata. Per le variabili indipendenti di scala e ordinale, gli algoritmi generano prima categorie utilizzando valori validi e poi decidono se unire la categoria mancante con la sua categoria (valida) più simile o mantenerla come categoria separata.
- Per CRT e QUEST i casi con valori variabili indipendenti mancanti sono esclusi dal processo di coltivazione dell'albero ma sono classificati utilizzando surrogati se i surrogati sono inclusi nel metodo. Se i valori nominali dell'utente - mancante sono trattati come mancanti, vengono gestiti anche in questo modo. Per ulteriori informazioni, consultare la sezione [“Surrogati”](#) a pagina 8.

#### Per Specificare nominale, Utente indipendente - Trattamento mancante

1. Nella finestra di dialogo Albero della decisione principale, selezionare almeno una variabile indipendente nominale.
2. Fare clic su **Opzioni**.
3. Fare clic sulla scheda **Valori mancanti**.

### Salvataggio delle informazioni sul modello

È possibile salvare le informazioni dal modello come variabili nel file dei dati di lavoro e si può anche salvare l'intero modello in formato XML (PMML) ad un file esterno.

Variabili salvate

**Numero nodo terminale.** Il nodo terminale a cui viene assegnato ogni caso. Il valore è il numero del nodo dell'albero.

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

**Probabilità previste.** La probabilità associata alla previsione del modello. Una variabile viene salvata per ogni categoria della variabile dipendente. Non disponibile per variabili dipendenti in scala.

**Assegnazione di esempio ( tirocinio / testing).** Per la validazione del campione suddiviso, questa variabile indica se un caso è stato utilizzato nel campione di formazione o testing. Il valore è di 1 per il campione di formazione e 0 per il campione di prova. Non disponibile a meno che non sia stata selezionata la validazione del campione split. Per ulteriori informazioni, consultare la sezione [“Convalida” a pagina 4.](#)

Esporta modello ad albero come XML

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

**Esempio di formazione.** Scrive il modello al file specificato. Per gli alberi convalidati a campione, questo è il modello per il campione formativo.

**Esempio di test.** Scrive il modello per il campione di test al file specificato. Non disponibile a meno che non sia stata selezionata la validazione del campione split.

## Output

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

### Visualizzazione struttura ad albero

È possibile controllare l'aspetto iniziale dell'albero o sopprimere completamente il display dell'albero.

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

**Visualizza.** Queste opzioni controllano l'aspetto iniziale del diagramma dell'albero nel Viewer. Tutti questi attributi possono essere modificati anche modificando l'albero generato.

- **Orientamento.** L'albero può essere visualizzato in alto con il nodo root in alto, da sinistra a destra o da destra a sinistra.
- **Contenuto del nodo.** I nodi possono visualizzare tabelle, grafici o entrambi. Per le variabili dipendenti categoriali, le tabelle visualizzano i conteggi di frequenza e le percentuali e i grafici sono grafici a barre. Per variabili dipendenti in scala, tabelle di visualizzazione medie, deviazioni standard, numero di casi e valori previsti e i grafici sono istogrammi.
- **Scala.** Per impostazione predefinita, gli alberi di grandi dimensioni vengono scaldati automaticamente in un tentativo di adattamento dell'albero alla pagina. È possibile specificare una percentuale di scala personalizzata fino al 200%.
- **Statistiche variabili indipendenti.** Per CHAID e CHAID esaustivo, le statistiche includono il valore  $F$  (per le variabili dipendenti in scala) o il valore di chi - quadrato (per le variabili dipendenti categoriali) nonché il valore di significatività e gradi di libertà. Per CRT, viene mostrato il valore di miglioramento. Per QUEST,  $F$ , valore di significatività e gradi di libertà sono mostrati per le variabili indipendenti di scala e ordinali; per le variabili indipendenti nominali, si mostrano chi - quadrato, valore di significatività e gradi di libertà.
- **Definizioni del nodo.** Le definizioni dei nodi visualizzano i valori della variabile indipendente utilizzati ad ogni nodo split.

**Albero in formato tabella.** Informazioni di riepilogo per ciascun nodo dell'albero, incluso numero di nodo principale, statistiche variabili indipendenti, valore variabile indipendente (s) per il nodo, media e deviazione standard per variabili dipendenti in scala, o conteggi e percentuali per variabili dipendenti categoriali.

Per controllare il Display Tree iniziale

1. Nella finestra di dialogo Albero della decisione principale, fare clic su **Output**.
2. Fare clic sulla scheda **Albero**.

## Statistiche

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

Modello

**Riepilogo.** Il riepilogo include il metodo utilizzato, le variabili incluse nel modello e le variabili specificate ma non incluse nel modello.

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

- Per le variabili dipendenti categoriali, la stima del rischio è la percentuale di casi erroneamente classificati dopo la regolazione per le probabilità precedenti e i costi di misclassificazione.
- Per le variabili dipendenti in scala, la stima del rischio è di varianza all'interno del nodo.

**Tabella di classificazione.** Per variabili dipendenti categoriali (nominali, ordinali), questa tabella mostra il numero di casi classificati correttamente e erroneamente per ogni categoria della variabile dipendente. Non disponibile per variabili dipendenti in scala.

**Costo, probabilità preventiva, punteggio e valori di profitto.** Per le variabili dipendenti categoriali, questa tabella mostra il costo, la probabilità preventiva, il punteggio e i valori di profitto utilizzati nell'analisi. Non disponibile per variabili dipendenti in scala.

Variabili indipendenti

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

**Surrogati per split.** Per i metodi di coltivazione CRT e QUEST, se il modello include surrogati, elenca i surrogati per ogni split nell'albero. Non disponibile per i metodi CHAID. Per ulteriori informazioni, consultare la sezione [“Surrogati” a pagina 8](#).

Prestazione del nodo

**Riepilogo.** Per le variabili dipendenti in scala, la tabella include il numero di nodo, il numero di casi e il valore medio della variabile dipendente. Per le variabili dipendenti categoriali con utili definiti, la tabella include il numero di nodo, il numero di casi, l'utile medio e i valori ROI (return on investment). Non disponibile per variabili dipendenti categoriali senza profitti definiti. Per ulteriori informazioni, consultare la sezione [“Profitti” a pagina 9](#).

**Per categoria di destinazione.** Per variabili dipendenti categoriali con categorie target definite, la tabella include il guadagno percentuale, la percentuale di risposta e la percentuale di indice (lift) per nodo o gruppo percentile. Viene prodotta una tabella separata per ogni categoria di destinazione. Non disponibile per variabili dipendenti in scala o variabili dipendenti categoriali senza categorie di destinazione definite. Per ulteriori informazioni, consultare la sezione [“Selezione Categorie” a pagina 4](#).

**Righe.** Le tabelle delle prestazioni del nodo possono visualizzare i risultati per nodi terminali, percentili o entrambi. Se si selezionano entrambi, vengono prodotte due tabelle per ogni categoria di destinazione. Le tabelle percentili visualizzano valori cumulativi per ogni percentile, in base all'ordine di ordinamento.

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

**Visualizza statistiche cumulative.** Per le tabelle dei nodi terminali, visualizza ulteriori colonne in ogni tabella con risultati cumulativi.

Per selezionare l'output di statistiche

1. Nella finestra di dialogo Albero della decisione principale, fare clic su **Output**.
2. Fare clic sulla scheda **Statistiche**.

## Grafici

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

**L'importanza variabile indipendente per il modello.** Grafico a barre di importanza del modello per variabile indipendente (predittore). Disponibile solo con il metodo di coltivazione CRT.

Prestazione del nodo

**Guadagno.** Guadagno è la percentuale di casi totali nella categoria target in ogni nodo, calcolata come:  $(\text{destinazione nodo } n / \text{destinazione totale } n) \times 100$ . Il grafico dei guadagni percentili cumulativi, calcolato come:  $(\text{target percentile cumulativo } n / \text{target totale } n) \times 100$ . Viene prodotto un grafico di linea separato per ogni categoria di destinazione. Disponibile solo per variabili dipendenti categoriali con categorie target definite. Per ulteriori informazioni, consultare la sezione [“Selezione Categorie”](#) a pagina 4.

I guadagni grafici compiono gli stessi valori che si vedrebbero nella colonna *Percentuale Gain* nei guadagni per la tabella dei percentili, che riporta anche i valori cumulativi.

**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 dell'indice è un grafico di linea dei valori di indice cumulativo dei percentili. Disponibile solo per variabili dipendenti categoriali. L'indice cumulativo del percentile è calcolato come:  $(\text{percentuale di risposta percentile cumulativo} / \text{risposta totale}) \times 100$ . Viene prodotto un grafico separato per ogni categoria di destinazione e le categorie di destinazione devono essere definite.

Il grafico dell'indice compila gli stessi valori che si vedrebbero nella colonna *Indice* nei guadagni per la tabella dei percentili.

**Risposta.** La percentuale di casi nel nodo nella categoria di destinazione specificata. Il grafico di risposta è un grafico a righe di risposta percentile cumulativo, calcolato come:  $(\text{target percentile cumulativo } n / \text{percentile cumulativo } n) \times 100$ . Disponibile solo per variabili dipendenti categoriali con categorie target definite.

Il grafico di risposta compila gli stessi valori che si vedrebbero nella colonna *Risposta* nei guadagni per la tabella dei percentili.

**Media.** Grafico di linea dei valori medi percentili cumulativi per la variabile dipendente. Disponibile solo per variabili dipendenti in scala.

**Utente medio.** Grafico di linea del profitto medio cumulativo. Disponibile solo per variabili dipendenti categoriali con utili definiti. Per ulteriori informazioni, consultare la sezione [“Profitti”](#) a pagina 9.

Il grafico di profitto medio compila gli stessi valori che si vedrebbero nella colonna *Profit* nel riepilogo dei guadagni per la tabella dei percentili.

**Torna sull'investimento (ROI).** Grafico di linea del ROI cumulativo (rendimento degli investimenti). Il ROI è calcolato come il rapporto dei profitti alle spese. Disponibile solo per variabili dipendenti categoriali con utili definiti.

Il grafico ROI compila gli stessi valori che si vedrebbero nella colonna *ROI* nel riepilogo dei guadagni per la tabella dei percentili.

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

Per selezionare Output del grafico

1. Nella finestra di dialogo Albero della decisione principale, fare clic su **Output**.
2. Fare clic sulla scheda **Pagine**.

## Regole di selezione e di valutazione

La scheda Regole fornisce la possibilità di generare regole di selezione o di classificazione / previsione sotto forma di sintassi del comando, SQL o semplice (plain English). È possibile visualizzare queste regole nel Viewer e / o salvare le regole in un file esterno.

**Sintassi.** Controlla la forma delle regole di selezione in entrambi gli output visualizzati nelle regole di Viewer e di selezione salvate in un file esterno.

- **IBM SPSS Statistics.** Lingua di sintassi del comando. Le regole sono espresse come una serie di comandi che definiscono una condizione di filtro che può essere utilizzata per selezionare sottoinsiemi di casi o come istruzioni COMPUTE che possono essere utilizzate per ottenere i casi.
- **SQL.** Le regole SQL standard vengono generate per selezionare o estrarre record da un database o assegnare valori a quei record. Le regole SQL generate non includono alcun nome di tabella o altre informazioni sull'origine dati.
- **Testo semplice.** Plain English pseudo - code. Le regole sono espresse come una serie di istruzioni logiche "se ... poi" che descrivono le classificazioni o le previsioni del modello per ogni nodo. Le regole in questo modulo possono utilizzare etichette variabili e valori variabili e valori di dati.

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

- **Assegnare i valori ai casi.** Le regole possono essere utilizzate per assegnare le previsioni del modello ai casi che soddisfano i criteri di appartenenza dei nodi. Viene generata una regola separata per ogni nodo che soddisfa i criteri di appartenenza del nodo.
- **Seleziona casi.** Le regole possono essere utilizzate per selezionare i casi che soddisfano i criteri di appartenenza dei nodi. Per le regole IBM SPSS Statistics e SQL viene generata una singola regola per selezionare tutti i casi che soddisfano i criteri di selezione.

**Include i surrogati in IBM SPSS Statistics e le regole SQL.** Per CRT e QUEST è possibile includere i predicatori surrogati dal modello nelle regole. Le regole che includono i surrogati possono essere piuttosto complesse. In generale, se si vuole solo ricavare informazioni concettuali sul proprio albero, escludere i surrogati. Se alcuni casi hanno dati incompleti variabili indipendenti (predictor) e si desidera regole che imitano il tuo albero, includere i surrogati. Per ulteriori informazioni, consultare la sezione [“Surrogati”](#) a pagina 8.

**Nodi.** Controlla l'ambito delle regole generate. Viene generata una regola separata per ogni nodo incluso nell'ambito.

- **Tutti i nodi terminali.** Genera regole per ogni nodo terminale.
- **I migliori nodi terminali.** Genera regole per i primi nodi terminali  $n$  in base ai valori dell'indice. Se il numero supera il numero di nodi terminali presenti nell'albero, le regole vengono generate per tutti i nodi terminali. (Vedi nota di seguito.)
- **I migliori nodi terminali fino a una determinata percentuale di casi.** Genera regole per i nodi terminali per la percentuale top  $n$  di casi in base ai valori dell'indice. (Vedi nota di seguito.)
- **I nodi Terminal il cui valore di indice incontra o supera un valore di cutoff.** Genera regole per tutti i nodi terminali con un valore di indice superiore o uguale al valore specificato. Un valore di indice superiore a 100 significa che la percentuale di casi nella categoria target in quel nodo supera la percentuale nel nodo root. (Vedi nota di seguito.)
- **Tutti i nodi.** Genera regole per tutti i nodi.

*Nota 1:* la selezione del nodo in base ai valori di indice è disponibile solo per variabili dipendenti categoriali con categorie target definite. Se sono state specificate più categorie di destinazione, viene generata una serie separata di regole per ogni categoria di destinazione.

*Nota 2:* per IBM SPSS Statistics e le regole SQL per la selezione dei casi (non regole per l'assegnazione dei valori), **Tutti i nodi** e **Tutti i nodi terminali** genereranno efficacemente una regola che seleziona tutti i casi utilizzati nell'analisi.

**Le regole di esportazione in un file.** Salva le regole in un file di testo esterno.

È inoltre possibile generare e salvare le regole di selezione o di punteggio in modo interattivo, in base ai nodi selezionati nel modello di albero finale. Per ulteriori informazioni, consultare la sezione [“Regole di selezione e di valutazione dei casi”](#) a pagina 19.

*Nota:* se si applicano regole sotto forma di sintassi del comando a un altro file di dati, quel file di dati deve contenere variabili con gli stessi nomi delle variabili indipendenti incluse nel modello finale, misurati nella stessa metrica, con gli stessi valori mancanti definiti dall'utente (se presenti).

Per specificare le regole di selezione o di calcolo

1. Nella finestra di dialogo Albero della decisione principale, fare clic su **Output**.
2. Fare clic sulla scheda **Regole**.

## Editor degli alberi

---

Con l'Tree Editor è possibile:

- Nascondi e mostra rami di albero selezionati.
- Visualizzazione del controllo del contenuto del nodo, delle statistiche visualizzate al nodo splits e di altre informazioni.
- Modificare i colori del nodo, dello sfondo, del bordo, del grafico e dei font.
- Cambiare stile e dimensione del carattere.
- Modifica l'allineamento dell'albero.
- Selezionare sottoserie di casi per ulteriori analisi in base ai nodi selezionati.
- Creare e salvare le regole per la selezione o il punteggio dei casi in base ai nodi selezionati.

Per modificare un modello di albero:

1. Fare doppio clic sul modello di albero nella finestra del Viewer.  
oppure
2. Dal menu Modifica o dal menu a comparsa visualizzato facendo clic con il tasto destro del mouse scegliere:

### **Modifica**

Nascondi e mostra nodi

Per nascondere (collapse) tutti i nodi child in un ramo sotto un nodo principale:

1. Fare clic sul segno meno (-) nella piccola casella sotto l'angolo inferiore destro del nodo principale.  
Tutti i nodi sotto il nodo principale su quel ramo saranno nascosti.

Per mostrare (espandere) i nodi child in un ramo sotto un nodo principale:

2. Fare clic sul segno più (+) nella piccola casella sotto l'angolo inferiore destro del nodo principale.

*Nota:* Hiding i nodi child su una diramazione non è lo stesso di prunare un albero. Se si desidera un albero prunato, è necessario richiedere il pruning prima di creare l'albero, e i rami prati non sono inclusi nell'albero finale. Per ulteriori informazioni, consultare la sezione [“Alberi Di Pruning”](#) a pagina 7.

Selezione Nodi multipli

È possibile selezionare i casi, generare regole di scoring e selezioni ed eseguire altre azioni in base al nodo attualmente selezionato (s). Per selezionare più nodi:

1. Fare clic su un nodo che si desidera selezionare.
2. Ctrl - clicca sugli altri nodi che si desidera selezionare.

È possibile selezionare più nodi di sibling e / o nodi principali in un ramo e nodi child in un altro ramo. Non è possibile, tuttavia, utilizzare più selezioni su un nodo principale e un bambino / discendente dello stesso ramo del nodo.

## Lavorare con Large Trees

I modelli ad albero possono talvolta contenere tanti nodi e rami che è difficile o impossibile visualizzare l'intero albero a grandezza completa. Ci sono una serie di funzioni che potreste trovare utili quando si lavora con alberi di grandi dimensioni:

- **mappa struttura.** È possibile utilizzare la mappa dell'albero, una versione molto più piccola, semplificata dell'albero, per navigare sull'albero e selezionare i nodi. Per ulteriori informazioni, consultare la sezione [“Mappa dell'albero”](#) a pagina 17.
- **Scaling.** È possibile zoomare e ingrandire modificando la percentuale di scala per il display dell'albero. Per ulteriori informazioni, consultare la sezione [“Scaling dell'albero Display”](#) a pagina 17.
- **Visualizzazione nodo e ramo.** È possibile rendere più compatto un albero visualizzando solo tabelle o solo grafici nei nodi e / o sopprimere la visualizzazione delle etichette dei nodi o delle informazioni variabili indipendenti. Per ulteriori informazioni, consultare la sezione [“Controllo Informazioni Visualizzate nell'Albero”](#) a pagina 18.

### Mappa dell'albero

La mappa dell'albero fornisce una vista compatta e semplificata dell'albero che è possibile utilizzare per navigare sull'albero e selezionare i nodi.

Per utilizzare la finestra della mappa dell'albero:

1. Dai menu dell'Albero Editor scegliere:

#### **Visualizza > Mappa struttura**

- Il nodo attualmente selezionato è evidenziato sia nell'Albero Model Editor che nella finestra della mappa dell'albero.
- La parte dell'albero che è attualmente nella zona della vista Editor Model Editor è indicata con un rettangolo rosso nella mappa dell'albero. Fare clic con il tasto destro del mouse e trascinare il rettangolo per modificare la sezione dell'albero visualizzato nell'area della vista.
- Se si seleziona un nodo nella mappa dell'albero che non è attualmente nella zona della vista Albero dell'Albero, i turni di visualizzazione per includere il nodo selezionato.
- La selezione dei nodi multipli funziona la stessa nella mappa dell'albero come nell'Albero dell'Albero: Ctrl - clicca per selezionare più nodi. Non è possibile utilizzare più selezioni su un nodo principale e un bambino / discendente dello stesso ramo del nodo.

### Scaling dell'albero Display

Per impostazione predefinita, gli alberi vengono scaldati automaticamente per adattarsi alla finestra del Viewer, che può risultare in alcuni alberi che inizialmente sono molto difficili da leggere. È possibile selezionare un'impostazione di scala preimpostata o inserire il proprio valore di scala personalizzato compreso tra 5% e 200%.

Per modificare la scala dell'albero:

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

oppure

2. Dai menu dell'Albero Editor scegliere:

#### **Visualizza > Scala ...**

È inoltre possibile specificare un valore di scala prima di creare il modello di albero. Per ulteriori informazioni, consultare la sezione [“Output”](#) a pagina 12.

## Finestra Riepilogo Nodi

La finestra di riepilogo del nodo fornisce una visione più ampia dei nodi selezionati. È inoltre possibile utilizzare la finestra di riepilogo per visualizzare, applicare o salvare le regole di selezione o di calcolo in base ai nodi selezionati.

- Utilizzare il menu **Visualizza** nella finestra di riepilogo del nodo per passare tra le viste di una tabella di riepilogo, di un grafico o di regole.
- Utilizzare il menu **Regole** nella finestra di riepilogo del nodo per selezionare il tipo di regole che si desidera visualizzare. Per ulteriori informazioni, consultare la sezione [“Regole di selezione e di valutazione dei casi” a pagina 19](#).
- Tutte le viste nella finestra di riepilogo del nodo riflettono un riepilogo combinato per tutti i nodi selezionati.

Per utilizzare la finestra di riepilogo del nodo:

1. Selezionare i nodi nell'Albero dell'Albero. Per selezionare più nodi, utilizzare Ctrl-click.
2. Dai menu, scegliere:

**Visualizza > Riepilogo**

## Controllo Informazioni Visualizzate nell'Albero

Il menu **Opzioni** nell'Albero dell'Albero consente di controllare la visualizzazione dei contenuti del nodo, dei nomi delle variabili indipendenti (predittore) e delle statistiche, delle definizioni dei nodi e di altre impostazioni. Molte di queste impostazioni possono essere controllate anche dalla barra degli strumenti.

## Modifica Colori dell'albero e font di testo

È possibile modificare i seguenti colori nell'albero:

- Bordo nodo, sfondo e colore del testo
- Colore del ramo e del testo del ramo
- Colore sfondo albero
- Colore di evidenziazione della categoria prevista (variabili dipendenti categoriali)
- Colori grafici del nodo

È anche possibile modificare il tipo di carattere, lo stile e la dimensione per tutto il testo dell'albero.

*Nota:* non è possibile modificare gli attributi del colore o del carattere per singoli nodi o rami. Le modifiche del colore si applicano a tutti gli elementi dello stesso tipo e le modifiche dei font (diverse dal colore) si applicano a tutti gli elementi grafici.

Per modificare i colori e gli attributi del font di testo:

1. Utilizzare la barra degli strumenti per modificare gli attributi del carattere per l'intero albero o i colori per diversi elementi dell'albero. (ToolTips descrivono ogni controllo sulla barra degli strumenti quando si mette il cursore del mouse sul controllo.)

oppure

2. Fare doppio clic ovunque nell'Albero dell'Albero per aprire la finestra **Proprietà** o dai menu scegliere:

**Visualizza > Proprietà**

3. Per il bordo, il ramo, lo sfondo del nodo, la categoria prevista e lo sfondo dell'albero, clicca sulla scheda **Colore**.
4. Per i colori e gli attributi dei font, fare clic sulla scheda **Testo**.
5. Per i colori del grafico dei nodi, fare clic sulla scheda **Carte del nodo**.

## Regole di selezione e di valutazione dei casi

È possibile utilizzare l'Albero dell'Albero per:

- Selezionare sottoserie di casi in base ai nodi selezionati. Per ulteriori informazioni, consultare la sezione [“Casi di filtraggio” a pagina 19](#).
- Generare regole di selezione dei casi o regole di punteggio nella sintassi del comando IBM SPSS Statistics o in formato SQL. Per ulteriori informazioni, consultare la sezione [“Salvataggio delle regole di selezione e di calcolo” a pagina 19](#).

È inoltre possibile salvare automaticamente le regole in base a vari criteri quando si esegue la procedura dell'Albero della decisione per creare il modello di albero. Per ulteriori informazioni, consultare la sezione [“Regole di selezione e di valutazione” a pagina 15](#).

### Casi di filtraggio

Se si desidera saperne di più sui casi in un determinato nodo o gruppo di nodi, è possibile selezionare un sottoinsieme di casi per ulteriori analisi in base ai nodi selezionati.

1. Selezionare i nodi nell'Albero dell'Albero. Per selezionare più nodi, utilizzare Ctrl-click.
2. Dai menu, scegliere:

**Regole > Casi filtro ...**

3. Inserire un nome per la variabile di filtro. I casi dai nodi selezionati riceveranno un valore di 1 per questa variabile. Tutti gli altri casi riceveranno un valore di 0 e saranno esclusi dalla successiva analisi fino a quando non si modifica lo stato del filtro.
4. Fare clic su **OK**.

### Salvataggio delle regole di selezione e di calcolo

È possibile salvare le regole di selezione dei casi o segnare in un file esterno e poi applicare quelle regole ad una diversa origine dati. Le regole sono basate sui nodi selezionati nell'Albero dell'Albero.

**Sintassi.** Controlla la forma delle regole di selezione in entrambi gli output visualizzati nelle regole di Viewer e di selezione salvate in un file esterno.

- **IBM SPSS Statistics.** Lingua di sintassi del comando. Le regole sono espresse come una serie di comandi che definiscono una condizione di filtro che può essere utilizzata per selezionare sottoinsiemi di casi o come istruzioni COMPUTE che possono essere utilizzate per ottenere i casi.
- **SQL.** Le regole SQL standard vengono generate per selezionare / estrarre record da un database o assegnare valori a quei record. Le regole SQL generate non includono alcun nome di tabella o altre informazioni sull'origine dati.

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

- **Seleziona casi.** Le regole possono essere utilizzate per selezionare i casi che soddisfano i criteri di appartenenza dei nodi. Per le regole IBM SPSS Statistics e SQL viene generata una singola regola per selezionare tutti i casi che soddisfano i criteri di selezione.
- **Assegnare i valori ai casi.** Le regole possono essere utilizzate per assegnare le previsioni del modello ai casi che soddisfano i criteri di appartenenza dei nodi. Viene generata una regola separata per ogni nodo che soddisfa i criteri di appartenenza del nodo.

**Include surrogati.** Per CRT e QUEST è possibile includere i predicatori surrogati dal modello nelle regole. Le regole che includono i surrogati possono essere piuttosto complesse. In generale, se si vuole solo ricavare informazioni concettuali sul proprio albero, escludere i surrogati. Se alcuni casi hanno dati incompleti variabili indipendenti (predictor) e si desidera regole che imitano il tuo albero, includere i surrogati. Per ulteriori informazioni, consultare la sezione [“Surrogati” a pagina 8](#).

Per salvare le regole di selezione dei casi o di punteggio:

1. Selezionare i nodi nell'Albero dell'Albero. Per selezionare più nodi, utilizzare Ctrl-click.

2. Dai menu, scegliere:

**Regole > Esporta ...**

3. Selezionare il tipo di regole desiderate ed inserire un nomefile.

*Nota:* se si applicano regole sotto forma di sintassi del comando a un altro file di dati, quel file di dati deve contenere variabili con gli stessi nomi delle variabili indipendenti incluse nel modello finale, misurati nella stessa metrica, con gli stessi valori mancanti definiti dall'utente (se presenti).

## Informazioni particolari

---

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

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

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

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

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

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

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

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

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

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

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

*IBM Director of Licensing*  
*IBM Corporation*

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

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

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

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

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

Qualsiasi affermazione relativa agli obiettivi e alla direzione futura di IBM è soggetta a modifica o revoca senza preavviso e concerne esclusivamente gli scopi dell'azienda.

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

#### LICENZA DI COPYRIGHT:

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

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

© Copyright IBM Corp. 2021. Le porzioni di questo codice derivano da IBM Corp. Programmi Di Esempio.

© Copyright IBM Corp. 1989 - 2021. Tutti i diritti riservati.

## Marchi

---

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

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

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

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

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

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

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



# Indice analitico

## A

### Alberi

- attributi di testo [18](#)
- caratteri [18](#)
- colori [18](#)
- colori grafici del nodo [18](#)
- contenuto albero in una tabella [12](#)
- controllo dimensione del nodo [5](#)
- controllo visualizzazione dell'albero [12](#), [18](#)
- Convalida del campione suddiviso [4](#)
- Convalida incrociata [4](#)
- costi di errata classificazione [8](#)
- Criteri di crescita CHAID [5](#)
- Generazione regole [15](#), [19](#)
- grafici [14](#)
- importanza predittore [13](#)
- Intervalli per variabili indipendenti di scala [6](#)
- lavorare con alberi di grandi dimensioni [17](#)
- limitare numero di livelli [5](#)
- mappa ad albero [17](#)
- Metodo CRT [7](#)
- modifica [16](#)
- mostrare e nascondere le statistiche del ramo [12](#)
- nascondere rami e nodi [16](#)
- orientamento all'albero [12](#)
- Probabilità a priori [9](#)
- profitti [9](#)
- punteggi [10](#)
- salvataggio variabili del modello [11](#)
- selezione di più nodi [16](#)
- statistiche del nodo terminale [13](#)
- stime di rischio [13](#)
- tabella di errata classificazione [13](#)
- taglio [7](#)
- valori di indice [13](#)
- valori mancanti [11](#)
- visualizzazione dell'albero di scaling [17](#)

### alberi decisionali

- forzatura della prima variabile in modello [1](#)
- livello di misurazione [1](#)
- Metodo CHAID [1](#)
- Metodo CHAID esaustivo [1](#)
- Metodo CRT [1](#)
- Metodo QUEST [1](#), [7](#)

## C

### CHAID

- Intervalli per variabili indipendenti di scala [6](#)
- Numero massimo di iterazioni [5](#)
- Regolazione di Bonferroni [5](#)
- rispettare categorie fuse [5](#)
- splitting e criteri di fusione [5](#)

### comprima rami dell'albero [16](#)

### convalida

- Alberi [4](#)

### Convalida del campione suddiviso

- Alberi [4](#)

### Convalida incrociata

- Alberi [4](#)

### costi

- errata classificazione [8](#)

### CRT

- misure di impurità [7](#)
- taglio [7](#)

## E

### errata classificazione

- Alberi [13](#)
- costi [8](#)

## G

### Gini [7](#)

## I

### Impurità

- Alberi CRT [7](#)

## L

### livello di misurazione

- alberi decisionali [1](#)

### Livello di significatività per la suddivisione dei nodi [7](#)

## N

### nascondere nodi

- vs. pruning [7](#)

### nascondere rami di albero [16](#)

### nod

- selezione di più nodi ad albero [16](#)

### Numero di nodi

- risparmio come variabile dagli alberi di decisione [11](#)

## P

### peso dei casi

- pesi frazionali negli alberi di decisione [1](#)

### Probabilità prevista

- risparmio come variabile dagli alberi di decisione [11](#)

### profitti

- Alberi [9](#), [13](#)
- Probabilità a priori [9](#)

### punteggi

- Alberi [10](#)

## Q

QUEST

taglio [7](#)

## R

regole

creazione della selezione e della sintassi di valutazione  
degli alberi di decisione [15](#), [19](#)

## S

selezione di più nodi ad albero [16](#)

seme per numeri casuali

validazione albero decisionale [4](#)

sintassi

creazione della selezione e della sintassi di valutazione  
degli alberi di decisione [15](#), [19](#)

sintassi del comando

creazione della selezione e della sintassi di valutazione  
degli alberi di decisione [15](#), [19](#)

SQL

creazione della sintassi SQL per la selezione e il  
punteggio [15](#), [19](#)

stime di rischio

Alberi [13](#)

## T

taglio strutture ad albero delle decisioni

vs. nascondere nodi [7](#)

Twoing [7](#)

Twoing ordinato [7](#)

## V

valori attesi

risparmio come variabile dagli alberi di decisione [11](#)

valori di indice

Alberi [13](#)

valori mancanti

Alberi [11](#)



