

*IBM SPSS Decision Trees 26*

**IBM**

**Hinweis**

Vor Verwendung dieser Informationen und des darin beschriebenen Produkts sollten die Informationen unter „Bemerkungen“ auf Seite 19 gelesen werden.

**Produktinformation**

Diese Ausgabe bezieht sich auf Version 26, Release 0, Modifikation 0 von IBM SPSS Statistics und alle nachfolgenden Releases und Modifikationen, bis dieser Hinweis in einer Neuausgabe geändert wird.

---

# Inhaltsverzeichnis

|  |          |                             |           |
|--|----------|-----------------------------|-----------|
| <b>Entscheidungsbäume . . . . .</b>          | <b>1</b> | Ausgabe . . . . .           | 13        |
| Erstellen von Entscheidungsbäumen . . . . .  | 1        | <b>Bemerkungen. . . . .</b> | <b>19</b> |
| Auswählen von Kategorien . . . . .           | 4        | Marken. . . . .             | 21        |
| Validierung. . . . .                         | 5        | <b>Index . . . . .</b>      | <b>23</b> |
| Kriterien für den Aufbau des Baums . . . . . | 5        |                             |           |
| Optionen . . . . .                           | 9        |                             |           |
| Speichern von Modellinformationen . . . . .  | 13       |                             |           |



---

# Entscheidungsbäume

Die folgenden Entscheidungsbaumfunktionen sind in SPSS Statistics Professional Edition oder der Option "Decision Trees" enthalten.

---

## Erstellen von Entscheidungsbäumen

Mit der Prozedur "Entscheidungsbaum" wird ein baumbasiertes Klassifizierungsmodell erstellt. Die Fälle werden in Gruppen klassifiziert oder es werden Werte für eine abhängige Variable (Zielvariable) auf der Grundlage der Werte von unabhängigen Variablen (Prädiktorvariablen) vorhergesagt. Die Prozedur umfasst Validierungstools für die explorative und die bestätigende Klassifikationsanalyse.

Die Prozedur eignet sich für folgende Situationen:

**Segmentierung.** Ermitteln Sie Personen, die wahrscheinlich zu einer bestimmten Gruppe gehören.

**Schichtung.** Weisen Sie Fälle einer von mehreren Kategorien zu, z. B. Gruppen mit hohem, mittlerem oder niedrigem Risiko.

**Vorhersage.** Erstellen Sie Regeln und lassen Sie damit zukünftige Ereignisse voraussagen, z. B. die Wahrscheinlichkeit, dass eine Person mit dem Darlehen in Bezug gerät, oder den potenziellen Wiederverkaufswert eines Autos oder Hauses.

**Dimensionsreduktion und Variablenscreening.** Wählen Sie ein geeignetes Subset an Prädiktoren aus einer Vielzahl von Variablen aus und bauen Sie damit ein formales parametrisches Modell auf.

**Erkennen von Interaktionen.** Ermitteln Sie Beziehungen, die nur für bestimmte Untergruppen gelten, und halten Sie diese in einem formalen parametrischen Modell fest.

**Zusammenführung von Kategorien und Diskretisierung stetiger Variablen.** Nehmen Sie die Umcodierung der Prädiktorkategorien und der stetigen Variablen bei minimalem Datenverlust vor.

**Beispiel.** Eine Bank möchte die Kreditantragsteller danach kategorisieren, ob sie ein annehmbares Kreditrisiko darstellen oder nicht. Auf der Grundlage verschiedener Faktoren (z. B. bekanntes Kreditrating bisheriger Kunden) können Sie ein Modell aufbauen, mit dem Sie vorhersagen, ob zukünftige Kunden mit ihren Darlehen in Verzug geraten würden.

Eine baumbasierte Analyse bietet einige attraktive Möglichkeiten:

- Sie können homogene Gruppen mit hohem oder niedrigem Risiko erkennen.
- Regeln für Vorhersagen zu individuellen Fällen können leichter aufgestellt werden.

## Erläuterung der Daten

**Daten.** Die abhängigen und die unabhängigen Variablen können wie folgt gestaltet sein:

- *Nominal.* Eine Variable kann als nominal behandelt werden, wenn ihre Werte Kategorien darstellen, die sich nicht in eine natürliche Reihenfolge bringen lassen, z. B. die Firmenabteilung, in der eine Person arbeitet. Beispiele für nominale Variablen sind Region, Postleitzahl oder Religionszugehörigkeit.
- *Ordinal.* Eine Variable kann als ordinal behandelt werden, wenn ihre Werte für Kategorien stehen, die eine natürliche Reihenfolge aufweisen (z. B. Grad der Zufriedenheit mit Kategorien von sehr unzufrieden bis sehr zufrieden). Ordinale Variablen treten beispielsweise bei Einstellungsmessungen (Zufriedenheit oder Vertrauen) und bei Präferenzbeurteilungen auf.

- *Skala*. Eine Variable kann als metrisch (stetig) behandelt werden, wenn ihre Werte geordnete Kategorien mit einer sinnvollen Metrik darstellen, sodass man sinnvolle Aussagen über die Abstände zwischen den Werten machen kann. Metrische Variablen sind beispielsweise Alter (in Jahren) oder Einkommen (in Geldeinheiten).

**Häufigkeitsgewichtungen** Wenn die Gewichtung aktiv ist, werden die Häufigkeitsgewichtungen auf die nächstliegende Ganzzahl gerundet. Fälle mit einer Gewichtung unter 0,5 erhalten einen Gewichtungswert von 0 und werden daher aus der Analyse ausgeschlossen.

**Annahmen.** Bei dieser Prozedur wird angenommen, dass allen Analysevariablen das entsprechende Messniveau zugewiesen wurde. Bei einigen Funktionen wird vorausgesetzt, dass eine Wertbeschriftung für alle Werte der in der Analyse berücksichtigten abhängigen Variablen definiert wurde.

- **Messniveau.** Das Messniveau beeinflusst die Baumberechnungen. Sämtlichen Variablen sollte daher das geeignete Messniveau zugewiesen werden. Standardmäßig wird angenommen, dass numerische Variablen metrisch und Zeichenfolgevariablen nominal sind; dies spiegelt gegebenenfalls nicht das tatsächliche Messniveau wider. Der Variablentyp ist durch ein Symbol neben der jeweiligen Variablen in der Variablenliste gekennzeichnet.

Sie können das Messniveau für eine Variable vorübergehend ändern. Klicken Sie hierzu mit der rechten Maustaste in der Liste der Quellenvariablen auf die entsprechende Variable und wählen Sie das gewünschte Messniveau im Popup-Menü.

- **Wertbeschriftungen.** In den Dialogfeldern für diese Prozedur wird angenommen, dass entweder alle der nicht fehlenden Werte einer kategorialen (nominalen, ordinalen) abhängigen Variablen über definierte Wertbeschriftungen verfügen oder keiner dieser Werte. Einige Funktionen sind nicht verfügbar, wenn nicht mindestens zwei nicht fehlende Werte der kategorialen abhängigen Variablen Wertbeschriftungen aufweisen. Wenn für mindestens zwei nicht fehlende Werte Wertbeschriftungen definiert sind, werden alle Fälle mit anderen Werten, die keine Wertbeschriftungen aufweisen, aus der Analyse ausgeschlossen.

## So erhalten Sie Entscheidungsbäume

1. Wählen Sie in den Menüs Folgendes aus:  
**Analysieren > Klassifizieren > Baum...**
2. Wählen Sie eine abhängige Variable aus.
3. Wählen Sie mindestens eine unabhängige Variable aus.
4. Wählen Sie eine Aufbaumethode aus.

Die folgenden Optionen sind verfügbar:

- Ändern Sie das Messniveau für eine Variable in der Liste der Quellenvariablen.
- Lassen Sie die erste Variable aus der Liste der unabhängigen Variablen als erste Teilungsvariable aufnehmen.
- Wählen Sie eine Einflussvariable aus, mit der definiert wird, wie viel Einfluss ein Fall auf den Aufbauprozess des Baums hat. Fälle mit niedrigeren Einflusswerten wirken sich weniger stark aus, Fälle mit höheren Werten entsprechend stärker. Die Einflussvariablen müssen positiv sein.
- Validieren Sie den Baum.
- Passen Sie die Kriterien für den Aufbau des Baums an.
- Speichern Sie die Endknotennummern, die vorhergesagten Werte und die vorhergesagten Wahrscheinlichkeiten als Variablen.
- Speichern Sie das Modell im XML-Format (PMML).

## Felder mit unbekanntem Messniveau

Der Messniveau-Alert wird angezeigt, wenn das Messniveau für mindestens eine Variable (ein Feld) im Dataset unbekannt ist. Da sich das Messniveau auf die Berechnung der Ergebnisse für diese Prozedur auswirkt, müssen alle Variablen ein definiertes Messniveau aufweisen.

### Daten durchsuchen

Liest die Daten im aktiven Dataset und weist allen Feldern, deren Messniveau zurzeit nicht bekannt ist, das Standardmessniveau zu. Bei großen Datasets kann dieser Vorgang einige Zeit in Anspruch nehmen.

### Manuell zuweisen

Listet alle Felder mit einem unbekanntem Messniveau auf. Sie können diesen Feldern ein Messniveau zuweisen. Sie können das Messniveau auch im Teilfenster "Variablenliste" des Dateneditors zuweisen.

Da das Messniveau für diese Prozedur bedeutsam ist, können Sie diese Prozedur erst dann ausführen, wenn für alle Felder ein Messniveau definiert wurde.

### Ändern des Messniveaus

1. Klicken Sie mit der rechten Maustaste auf eine Variable in der Liste der Quellenvariablen.
2. Wählen Sie ein Messniveau im Popup-Menü aus.

Das Messniveau wird vorübergehend für die Dauer der Prozedur "Entscheidungsbaum" geändert.

### Aufbaumethoden

Die folgenden Aufbaumethoden sind verfügbar:

#### CHAID

Steht für "Chi-squared Automatic Interaction Detection", d. h. automatische Erkennung von Interaktionen mittels Chi-Quadrat-Tests. In jedem Schritt bestimmt das CHAID-Verfahren diejenige unabhängige Variable (Prädiktor), die den stärksten Zusammenhang mit der abhängigen Variablen aufweist. Die Kategorien der einzelnen Prädiktoren werden zusammengeführt, wenn sie im Hinblick auf die abhängige Variable nicht signifikant unterschiedlich sind.

#### Exhaustive CHAID

Eine Abwandlung von CHAID, die für jede Prädiktorvariable alle möglichen Aufteilungen untersucht.

**CRT** Steht für "Classification and Regression Trees", d. h. Klassifikations- und Regressionsbäume. CRT unterteilt die Daten in Segmente, die im Hinblick auf die abhängige Variable so homogen wie möglich sind. Ein Endknoten, in dem alle Fälle denselben Wert der abhängigen Variablen haben, ist ein homogener ("reiner") Knoten.

#### QUEST

Steht für Quick, Unbiased, Efficient Statistical Tree, d. h. schneller, unverzerrter, effizienter statistischer Baum. Dabei handelt es sich um ein schnelles Verfahren, das die in anderen Verfahren auftretende Verzerrung zugunsten von Prädiktoren mit vielen Kategorien vermeidet. QUEST kann nur dann gewählt werden, wenn die abhängige Variable nominal ist.

Jede Methode hat ihre Vorteile und Einschränkungen:

Tabelle 1. Merkmale der Aufbaumethode.

| Merkmal  | CHAID* | CRT | QUEST |
|--|--------|-----|-------|
| Chi-Quadrat-basiert**                                    | O      |     |       |
| Surrogate für unabhängige Variablen (Prädiktorvariablen) |        | O   | O     |
| Beschneiden des Baums                                    |        | O   | O     |

Tabelle 1. Merkmale der Aufbaumethode (Forts.).

| Merkmal                            | CHAID* | CRT | QUEST |
|------------------------------------|--------|-----|-------|
| Aufteilen mehrdimensionaler Knoten | O      |     |       |
| Aufteilen binärer Knoten           |        | O   | O     |
| Einflussvariablen                  | O      | O   |       |
| A-priori-Wahrscheinlichkeiten      |        | O   | O     |
| Fehlklassifizierungskosten         | O      | O   | O     |
| Schnelle Berechnung                | O      |     | O     |

\*Mit Exhaustive CHAID.

\*\*Bei QUEST wird auch ein Chi-Quadrat-Maß für nominale unabhängige Variablen verwendet.

## Auswählen von Kategorien

Bei kategorialen (nominalen, ordinalen) abhängigen Variablen stehen folgende Möglichkeiten zur Auswahl:

- Legen Sie Kategorien fest, die im Diagramm angezeigt werden sollen.
- Geben Sie die relevanten Zielkategorien an.

## Kategorien ein-/ausschließen

Sie können die Analyse auf bestimmte Kategorien der abhängigen Variablen einschränken.

- Fälle mit Werten der abhängigen Variablen in der Liste "Ausschließen" werden bei der Analyse nicht berücksichtigt.
- Bei nominalen abhängigen Variablen können auch benutzerdefiniert fehlende Kategorien in die Analyse aufgenommen werden. (Standardmäßig werden benutzerdefiniert fehlende Kategorien in der Liste "Ausschließen" aufgeführt.)

## Zielkategorien

Die ausgewählten (markierten) Kategorien werden als primär relevante Kategorien in der Analyse behandelt. Wenn Sie beispielsweise hauptsächlich die Personen ermitteln möchten, bei denen die Wahrscheinlichkeit groß ist, dass sie mit ihrem Darlehen in Verzug geraten, bestimmen Sie entsprechend die Kategorie für schlechtes Kreditrating als Zielkategorie.

- Es ist keine Standardzielkategorie festgelegt. Ist keine Kategorie ausgewählt, stehen einige Optionen für die Klassifikation sowie die Ausgabe im Zusammenhang mit dem Profit nicht zur Verfügung.
- Wenn mehrere Kategorien angegeben sind, werden separate Tabellen und Diagramme mit dem Profit in den einzelnen Zielkategorien erstellt.
- Die Kennzeichnung von einer oder mehreren Kategorien als Zielkategorien wirkt sich nicht auf das Baummodell, die Risikoschätzung und die Fehlklassifizierungsergebnisse aus.

## Kategorien

In diesem Dialogfeld sind definierte Wertbeschriftungen für die abhängige Variable erforderlich. Das Dialogfeld ist erst dann verfügbar, wenn mindestens zwei Werte der kategorialen abhängigen Variablen eine Wertbeschriftung besitzen.

## So schließen Sie Kategorien ein- oder aus und wählen Zielkategorien aus

1. Wählen Sie im Hauptdialogfeld "Entscheidungsbaum" eine kategoriale (nominale, ordinale) abhängige Variable mit mindestens zwei definierten Wertbeschriftungen aus.



2. Klicken Sie auf **Kategorien**.

## Validierung

Mit der Validierung stellen Sie fest, wie gut sich die Baumstruktur auf eine größere Gesamtheit verallgemeinern lässt. Es stehen zwei Validierungsmethoden zur Auswahl: Kreuzvalidierung und Split-Sample-Validierung.

### Kreuzvalidierung

Bei der Kreuzvalidierung wird die Stichprobe in mehrere Teilstichproben oder **Aufteilungen** gegliedert. Anschließend werden Baummodelle erzeugt; dabei werden nacheinander die Daten der einzelnen Stichproben ausgeschlossen. Der erste Baum beruht auf allen Fällen mit Ausnahme der Fälle in der ersten Stichprobenaufteilung, der zweite Baum auf allen Fällen mit Ausnahme der Fälle in der zweiten Stichprobenaufteilung usw. Bei jedem Baum wird jeweils das Fehlklassifizierungsrisiko geschätzt. Hierzu wird der Baum auf die Teilstichprobe angewendet, die beim Erstellen des Baums ausgeschlossen war.

**Wichtig:** Die Kreuzvalidierung ist für CRT- und Quest-Methoden nicht verfügbar, wenn das Beschneiden ausgewählt wird.

- Sie können bis zu 25 Stichprobenaufteilungen angeben. Je höher der Wert, desto weniger Fälle werden in den einzelnen Baummodellen ausgeschlossen.
- Bei der Kreuzvalidierung entsteht ein einziges, endgültiges Baummodell. Die kreuzvalidierte Risikoschätzung für den fertigen Baum wird als Durchschnitt des Risikos bei allen Bäumen berechnet.

### Split-Sample-Validierung

Bei der Split-Sample-Validierung wird das Modell mithilfe einer Trainingsstichprobe erzeugt und dann mit einer Holdout-Stichprobe überprüft.

- Sie können eine Trainingsstichprobe angeben (als Prozentsatz der gesamten Stichprobengröße) oder auch eine Variable, mit der die Stichprobe in Trainings- und Teststichproben aufgeteilt wird.
- Wenn Sie die Trainings- und Teststichproben mithilfe einer Variablen festlegen, werden Fälle mit dem Wert 1 für die Variable in die Trainingsstichprobe übernommen, alle anderen Fälle in die Teststichprobe. Die abhängige Variable, die Gewichtungvariable, die Einflussvariable sowie erzwungene unabhängige Variablen sind hier als Variable nicht zulässig.
- Die Ergebnisse können wahlweise für die Trainings- und Teststichproben oder auch nur für die Teststichprobe angezeigt werden.
- Bei kleinen Datendateien (Dateien mit nur wenigen Dateien) sollte die Split-Sample-Validierung nur nach sorgfältiger Erwägung verwendet werden. Kleine Trainingsstichproben können zu mangelhaften Modellen führen, weil einige Kategorien unter Umständen nicht genügend Fälle enthalten, damit der Baum ordnungsgemäß wachsen kann.

## Validieren eines Entscheidungsbaums

1. Klicken Sie im Hauptdialog "Entscheidungsbaum" auf **Validierung**.
2. Wählen Sie **Kreuzvalidierung** oder **Split-Sample-Validierung**.

**Anmerkung:** Bei beiden Validierungsmethoden werden die Fälle nach dem Zufallsprinzip zu den Stichprobengruppen zugewiesen. Sollen genau dieselben Ergebnisse in einer späteren Analyse reproduziert werden, bestimmen Sie den Startwert für die Zufallszahlen (Menü "Transformieren", "Zufallszahlengeneratoren"), bevor Sie die Analyse erstmalig ausführen, und geben Sie dann diesen Startwert für die Zufallszahlen bei der späteren Analyse ein.

## Kriterien für den Aufbau des Baums

Die verfügbaren Aufbaukriterien können von der Aufbaumethode und/oder dem Messniveau der abhängigen Variablen abhängen.

## Aufbaubegrenzungen

Im Dialogfeld "Aufbaubegrenzungen" können Sie die Anzahl der Ebenen im Baum einschränken und die Mindestanzahl der Fälle für über- und untergeordnete Knoten steuern.

### Maximale Baumtiefe

Steuert die maximale Anzahl der Aufbauebenen unterhalb des Stammknotens. Mit der Einstellung **Automatisch** wird der Baum auf drei (CHAID und Exhaustive CHAID) bzw. fünf Ebenen unterhalb des Stammknotens (CRT und QUEST) begrenzt.

### Minimale Anzahl von Fällen

Steuert die Mindestanzahl der Fälle für die Knoten. Knoten, die diese Kriterien nicht erfüllen, werden nicht aufgeteilt.

- Wenn Sie die Mindestwerte anheben, entstehen in der Regel Bäume mit weniger Knoten.
- Werden die Mindestwerte gesenkt, entstehen Bäume mit mehr Knoten.

Bei Datendateien mit nur wenigen Fällen führen die Standardwerte von 100 Fällen für übergeordnete Knoten und 50 Fällen für untergeordnete Knoten unter Umständen dazu, dass der resultierende Baum keine Knoten unterhalb des Stammknotens erhält. In dieser Situation sollten Sie die Mindestwerte verringern, um so aussagekräftigere Ergebnisse zu erzielen.

## So legen Sie die Aufbaubegrenzungen fest

1. Klicken Sie im Hauptdialogfeld "Entscheidungsbaum" auf **Aufbaubegrenzungen**.

## CHAID-Kriterien

Bei den Methoden CHAID und Exhaustive CHAID können Sie Folgendes steuern:

### Signifikanzniveau für

Legen Sie den Signifikanzwert für das Aufteilen von Knoten und das Zusammenführen von Kategorien fest. Bei beiden Kriterien liegt das Standardsignifikanzniveau bei 0,05.

#### Aufteilen von Knoten

Der Wert muss größer als 0 und kleiner als 1 sein. Niedrigere Werte führen in der Regel zu Bäumen mit weniger Knoten.

#### Zusammenführen von Kategorien

Der Wert muss größer als 0 und kleiner-gleich 1 sein. Wenn ein Zusammenführen der Kategorien unterbunden werden soll, legen Sie den Wert 1 fest. Bei einer metrischen unabhängigen Variablen bedeutet dies, dass die Anzahl der Kategorien für die Variable im fertigen Baum der angegebenen Anzahl an Intervallen entspricht (Standardwert: 10). Weitere Informationen finden Sie in „Metrische Intervalle für die CHAID-Analyse“ auf Seite 7.

## Chi-Quadrat-Statistik

Bei ordinalen abhängigen Variablen wird der Chi-Quadrat-Wert, mit dem das Aufteilen von Knoten und das Zusammenführen von Kategorien bestimmt wird, mithilfe der Likelihood-Quotienten-Methode berechnet. Bei nominalen abhängigen Variablen können Sie die Methode auswählen:

### Pearson

Diese Methode liefert schnellere Berechnungen, sollte bei kleineren Stichproben jedoch nur nach sorgfältiger Erwägung verwendet werden. Dies ist die Standardmethode.

### Likelihood-Quotient

Diese Methode ist stabiler als die Pearson-Methode; die Berechnungen nehmen jedoch mehr Zeit in Anspruch. Diese Methode eignet sich ideal für kleine Stichproben.

## Modellschätzung

Bei nominalen und ordinalen abhängigen Variablen können Sie Folgendes festlegen:

### Maximale Anzahl an Iterationen

Der Standardwert ist 100. Wenn der Baum nicht mehr weiter aufgebaut wird, weil die maximale Anzahl von Iterationen erreicht ist, können Sie den Maximalwert erhöhen oder auch Kriterien ändern, die den Aufbau des Baums steuern.

### **Mindeständerung bei den erwarteten Zellenhäufigkeiten**

Der Wert muss größer als 0 und kleiner als 1 sein. Der Standardwert ist 0,05. Bei niedrigeren Werten entstehen Bäume mit weniger Knoten.

### **Signifikanzwerte mit der Bonferroni-Methode anpassen**

Bei Mehrfachvergleichen werden die Signifikanzwerte für die Zusammenführungs- und Aufteilungskriterien mithilfe der Bonferroni-Methode angepasst. Dies ist die Standardeinstellung.

### **Erneute Aufteilung verbundener Kategorien innerhalb eines Knotens zulassen**

Sofern Sie das Zusammenführen von Kategorien nicht explizit unterbinden, werden Kategorien mit unabhängigen Variablen (Prädiktorvariablen) nach Möglichkeit zusammengeführt, um so den einfachsten Baum zu bilden, der das Modell beschreibt. Bei dieser Option können zusammengeführte Kategorien eigenständig durch die Prozedur erneut aufgeteilt werden, wenn hierdurch eine bessere Lösung entstünde.

### **So legen Sie die CHAID-Kriterien fest:**

1. Wählen Sie im Hauptdialogfeld "Entscheidungsbaum" **CHAID** oder **Exhaustive CHAID** als Aufbau- methode aus.
2. Klicken Sie auf **CHAID**.

**Metrische Intervalle für die CHAID-Analyse:** Bei der CHAID-Analyse werden metrische unabhängige Variablen (Prädiktorvariablen) vor der Analyse stets in diskrete Gruppen eingeteilt (z. B. 0-10, 11-20, 21-30 usw.). Sie können die anfängliche und maximale Anzahl der Gruppen steuern (unter Umständen werden aufeinander folgende Gruppen nach der ursprünglichen Aufteilung jedoch wieder zusammengeführt):

#### **Feste Zahl**

Alle metrischen unabhängigen Variablen werden zunächst in dieselbe Anzahl an Gruppen eingeteilt. Der Standardwert ist 10.

#### **Benutzerdefiniert**

Jede metrische unabhängige Variable wird zunächst in die Anzahl der Gruppen eingeteilt, die für die betreffende Variable angegeben sind.

### **So legen Sie die Intervalle für metrische unabhängige Variablen fest:**

1. Wählen Sie im Hauptdialogfeld "Entscheidungsbaum" mindestens eine metrische unabhängige Variable aus.
2. Wählen Sie als **Aufbaumethode** die Option **CHAID** oder **Exhaustive CHAID** aus.
3. Klicken Sie auf **Intervalle**.

Bei der CRT- und QUEST-Analyse werden nur binäre Aufteilungen verwendet und die metrischen und ordinalen unabhängigen Variablen werden auf dieselbe Weise behandelt. Es ist also nicht möglich, eine Intervallanzahl für die metrischen unabhängigen Variablen festzulegen.

### **CRT-Kriterien**

Bei der CRT-Aufbaumethode wird die Homogenität innerhalb der Knoten angestrebt. Das Ausmaß, in dem ein Knoten von einem homogenen Subset von Fällen abweicht, ist ein Hinweis auf **Unreinheit**. Beispiel: Ein Endknoten, in dem alle Fälle denselben Wert für die abhängige Variable aufweisen, ist ein homogener Knoten. Eine weitere Aufteilung ist nicht nötig, weil der Knoten bereits "rein" ist.

Sie können die Methode zum Messen der Unreinheit bestimmen und auch den Rückgang in der Unreinheit angeben, der mindestens erreicht werden muss, damit die Knoten aufgeteilt werden.

#### **Unreinheitsmaß**

Bei metrischen abhängigen Variablen wird das LSD-Unreinheitsmaß (Least-Squared Deviation, kleinste quadratische Abweichung) verwendet. Dieser Wert wird als Varianz innerhalb der Kno-

ten berechnet und gegebenenfalls gemäß der Häufigkeitsgewichtungen oder der Einflusswerte angepasst. Bei kategorialen (nominalen, ordinalen) abhängigen Variablen stehen die folgenden Unreinheitsmaße zur Auswahl:

**Gini** Die Aufteilungen maximieren die Homogenität der untergeordneten Knoten im Hinblick auf den Wert der abhängigen Variable. Das Gini-Maß beruht auf den quadratischen Wahrscheinlichkeiten für die Zugehörigkeit zu einer Kategorie der abhängigen Variable. Der Mindestwert (Null) wird erreicht, sobald alle Fälle in einem Knoten in eine einzige Kategorie fallen. Dies ist das Standardmaß.

**Twoing**

Die Kategorien der abhängigen Variablen werden in zwei Unterklassen gruppiert. Die Aufteilungen bewirken die bestmögliche Trennung der beiden Gruppen.

**Ordinales Twoing**

Dieses Maß entspricht weitgehend dem Twoing, mit der Ausnahme, dass nur nebeneinander liegende Kategorien gruppiert werden können. Dieses Maß steht nur bei ordinalen abhängigen Variablen zur Verfügung.

**Mindeständerung bei der Verbesserung**

Dies ist der mindestens erforderliche Rückgang der Unreinheit für das Aufteilen eines Knotens. Der Standardwert lautet 0.0001. Bei höheren Werten entstehen Bäume mit weniger Knoten.

**So legen Sie die CRT-Kriterien fest**

1. Wählen Sie als **Aufbaumethode** die Option **CRT** aus.
2. Klicken Sie auf **CRT**.

**QUEST-Kriterien**

Bei der QUEST-Methode können Sie das Signifikanzniveau für das Aufteilen von Knoten festlegen. Die Knoten können nur dann mit einer unabhängigen Variablen aufgeteilt werden, wenn das Signifikanzniveau kleiner-gleich dem angegebenen Wert ist. Der Wert muss größer als 0 und kleiner als 1 sein. Der Standardwert ist 0,05. Bei kleineren Werten werden mehr unabhängige Variablen aus dem endgültigen Modell ausgeschlossen.

**So legen Sie die QUEST-Kriterien fest**

1. Wählen Sie im Hauptdialogfeld "Entscheidungsbaum" eine nominale abhängige Variable aus.
2. Wählen Sie als **Aufbaumethode** die Option **QUEST** aus.
3. Klicken Sie auf **QUEST**.

**Beschneiden von Bäumen**

Bei der CRT- und der QUEST-Methode können Sie ein Überfüllen des Modells vermeiden, indem Sie den Baum **beschneiden**: Der Baum wächst, bis die Kriterien für das Anhalten erfüllt sind. Anschließend wird der Baum automatisch gemäß der angegebenen maximalen Risikodifferenz auf den kleinsten untergeordneten Baum getrimmt. Der Risikowert wird in Standardfehlern ausgedrückt. Der Standardwert ist 1. Der Wert muss positiv oder gleich Null sein. Um den untergeordneten Baum mit dem geringstmöglichen Risiko zu erzielen, geben Sie den Wert 0 an.

**Wichtig:** Die Kreuzvalidierung ist für CRT- und Quest-Methoden nicht verfügbar, wenn das Beschneiden ausgewählt wird.

**Beschneiden eines Baumes**

1. Wählen Sie im Hauptdialogfeld "Entscheidungsbaum" als **Aufbaumethode** die Option **CRT** oder **QUEST** aus.
2. Klicken Sie auf **Beschneiden**.

## Beschneiden im Unterschied zum Ausblenden von Knoten

Bei einem beschnittenen Baum sind alle Knoten, die aus dem Baum herausgeschnitten wurden, im endgültigen Baum nicht mehr verfügbar. Sie können zwar ausgewählte untergeordnete Knoten im fertigen Baum interaktiv ein- und ausblenden; es ist jedoch nicht möglich, Knoten anzeigen zu lassen, die beim Erstellen des Baums beschnitten wurden.

## Surrogate

Bei CRT und QUEST können **Surrogate** für unabhängige Variablen (Prädiktorvariablen) verwendet werden. In Situationen, in denen der Wert für die betreffende Variable fehlt, werden andere unabhängige Variablen, die einen hohen Grad an Zusammenhang mit der ursprünglichen Variable besitzen, zur Klassifizierung herangezogen. Diese alternativen Prädiktoren werden als Surrogate bezeichnet. Sie können die maximal zulässige Anzahl an Surrogaten für das Modell festlegen.

- Standardmäßig ist die maximale Anzahl an Surrogaten um 1 kleiner als die Anzahl der unabhängigen Variablen. Für eine unabhängige Variable kann also jede andere unabhängige Variable als Surrogat verwendet werden.
- Sollen keine Surrogate im Modell verwendet werden, geben Sie den Wert 0 als Anzahl der Surrogate an.

## So legen Sie Surrogate fest

1. Wählen Sie im Hauptdialogfeld "Entscheidungsbaum" als **Aufbaumethode** die Option **CRT** oder **QUEST** aus.
2. Klicken Sie auf **Surrogate**.

## Optionen

Die tatsächlich verfügbaren Optionen sind abhängig von der Aufbaumethode, dem Messniveau der abhängigen Variablen und/oder dem Vorhandensein definierter Wertbeschriftungen für die Werte der abhängigen Variable.

## Fehlklassifizierungskosten

Bei kategorialen (nominalen, ordinalen) abhängigen Variablen können Sie mit den Fehlklassifizierungskosten die relative Strafe für die fehlerhafte Klassifizierung angeben. Beispiel:

- Die Kosten, wenn einem kreditwürdigen Kunden ein Darlehen verweigert wird, unterscheiden sich in der Regel von den Kosten, wenn ein Kunde ein Darlehen erhält und dann damit in Verzug gerät.
- Die Kosten für die Fehlklassifizierung einer Person mit einem hohen Risiko für Herzerkrankungen als Person mit niedrigem Risiko sind wahrscheinlich deutlich höher, als wenn eine Person mit niedrigem Risiko fälschlicherweise mit einem hohen Risiko klassifiziert würde.
- Die Kosten für den Versand einer Werbesendung an eine Person, die wahrscheinlich nicht reagieren wird, sind relativ gering; die Kosten, wenn die Werbesendung nicht an eine Person geht, die wahrscheinlich reagiert hätte, sind dagegen deutlich höher (was den entgangenen Umsatz angeht).

**Anmerkung:** Dieses Dialogfeld "Fehlklassifizierungskosten" ist erst dann verfügbar, wenn mindestens zwei Werte der kategorialen abhängigen Variablen eine Wertbeschriftung besitzen.

## So legen Sie die Fehlklassifizierungskosten fest

1. Wählen Sie im Hauptdialogfeld "Entscheidungsbaum" eine kategoriale (nominale, ordinale) abhängige Variable mit mindestens zwei definierten Wertbeschriftungen aus.
2. Klicken Sie auf **Fehlklassifizierungskosten**.
3. Klicken Sie auf **Benutzerdefiniert**.
4. Geben Sie mindestens einen Wert für die Fehlklassifizierungskosten in das Raster "Vorhergesagte Kategorie" ein. Die Werte müssen positiv oder gleich Null sein. (Richtige Klassifizierungen, auf der Diagonalen dargestellt, sind stets gleich 0.)

## Füllmatrix

Häufig sollen die Kosten symmetrisch sein: Die Kosten für die Fehlklassifizierung von A als B sind genauso hoch wie die Kosten für die Fehlklassifizierung von B als A. Die folgenden Steuer-elemente erleichtern das Anlegen einer symmetrischen Kostenmatrix:

### Unteres Dreieck duplizieren

Kopiert Werte aus dem unteren Dreieck der Matrix (unterhalb der Diagonalen) in die entsprechenden Zellen oberhalb des Dreiecks.

### Oberes Dreieck duplizieren

Kopiert Werte aus dem oberen Dreieck der Matrix (oberhalb der Diagonalen) in die entsprechenden Zellen unterhalb des Dreiecks.

### Durchschnittliche Zellenwerte verwenden

Für jede Zelle in beiden Hälften der Matrix wird der Durchschnitt aus den beiden Werten (im oberen und unteren Dreieck) gebildet und anstelle der ursprünglichen beiden Werte eingesetzt. Beispiel: Die Fehlklassifizierung von A als B verursacht Kosten in Höhe von 1 und die Kosten für die Fehlklassifizierung von B als A betragen 3. Beide Werte werden somit durch den Durchschnitt  $(1+3)/2 = 2$  ersetzt.

## Profite

Bei kategorialen abhängigen Variablen können Sie den verschiedenen Ebenen jeweils Werte für Verkaufserlöse und Aufwendungen zuweisen.

- Der Profit ergibt sich aus der Berechnung Verkaufserlöse minus Aufwendungen.
- Die Profitwerte beeinflussen die Werte für den durchschnittlichen Profit und den Anlageertrag (ROI) in den Gewinntabellen. Die grundlegende Baummodellstruktur bleibt unverändert.
- Die Werte für Verkaufserlöse und Aufwendungen müssen numerisch sein und müssen für alle im Raster angezeigten Kategorien der abhängigen Variablen festgelegt werden.

**Anmerkung:** In diesem Dialogfeld sind definierte Wertbeschriftungen für die abhängige Variable erforderlich. Das Dialogfeld ist erst dann verfügbar, wenn mindestens zwei Werte der kategorialen abhängigen Variablen eine Wertbeschriftung besitzen.

## So geben Sie die Gewinne an

1. Wählen Sie im Hauptdialogfeld "Entscheidungsbaum" eine kategoriale (nominale, ordinale) abhängige Variable mit mindestens zwei definierten Wertbeschriftungen aus.
2. Klicken Sie auf **Gewinne**.
3. Klicken Sie auf **Benutzerdefiniert**.
4. Geben Sie die Werte für Verkaufserlöse und Aufwendungen für alle im Raster aufgeführten Kategorien der abhängigen Variablen ein.

## A-priori-Wahrscheinlichkeiten

Bei CRT- und QUEST-Bäumen mit kategorialen abhängigen Variablen können Sie A-priori-Wahrscheinlichkeiten für die Gruppenzugehörigkeit angeben. **A-priori-Wahrscheinlichkeiten** sind eine Schätzung der gesamten relativen Häufigkeit für jede Kategorie der abhängigen Variable, die aufgestellt wird, noch bevor die Werte der unabhängigen Variablen (Prädiktorvariablen) bekannt sind. Mithilfe von A-priori-Wahrscheinlichkeiten können Sie den Aufbau des Baums durch Daten in der Stichprobe korrigieren, die nicht repräsentativ für die Gesamtheit als Ganzes sind.

### Aus Trainingsstichprobe übernehmen (empirische A-priori-Wahrscheinlichkeiten)

Aktivieren Sie diese Einstellung, wenn die Verteilung der Variablenwerte in der Datendatei repräsentativ für die Verteilung in der Gesamtheit ist. Bei der Split-Sample-Validierung wird die Verteilung der Fälle in der Trainingsstichprobe herangezogen.

**Anmerkung:** Bei der Split-Sample-Validierung werden die Fälle nach dem Zufallsprinzip in die Trainingsstichprobe aufgenommen. Die eigentliche Verteilung der Fälle in der Trainingsstichprobe ist daher im Voraus nicht bekannt. Weitere Informationen finden Sie in „Validierung“ auf Seite 5.

### In allen Kategorien gleich

Aktivieren Sie diese Einstellung, wenn die Kategorien der abhängigen Variablen in der Gesamtheit gleichmäßig repräsentiert sind. Beispiel: Es liegen vier Kategorien vor und auf jede Kategorie entfallen etwa 25 % der Fälle.

### Benutzerdefiniert

Geben Sie je einen positiven Wert (oder den Wert 0) für jede im Raster aufgeführte Kategorie der abhängigen Variablen ein. Die Werte können Anteile, Prozentsätze oder Häufigkeitszähler umfassen oder auch andere Werte, die die Verteilung der Werte in den Kategorien wiedergeben.

### A-priori-Wahrscheinlichkeiten anhand der Fehlklassifizierungskosten korrigieren

Wenn Sie benutzerdefinierte Fehlklassifizierungskosten definieren, können Sie die A-priori-Wahrscheinlichkeiten anhand dieser Kosten anpassen. Weitere Informationen finden Sie in „Fehlklassifizierungskosten“ auf Seite 9.

**Anmerkung:** In diesem Dialogfeld sind definierte Wertbeschriftungen für die abhängige Variable erforderlich. Das Dialogfeld ist erst dann verfügbar, wenn mindestens zwei Werte der kategorialen abhängigen Variablen eine Wertbeschriftung besitzen.

### So legen Sie A-priori-Wahrscheinlichkeiten fest:

1. Wählen Sie im Hauptdialogfeld "Entscheidungsbaum" eine kategoriale (nominale, ordinale) abhängige Variable mit mindestens zwei definierten Wertbeschriftungen aus.
2. Wählen Sie als **Aufbaumethode** die Option **CRT** oder **QUEST** aus.
3. Klicken Sie auf **A-priori-Wahrscheinlichkeiten**.

### Scores

Bei CHAID und Exhaustive CHAID mit einer ordinalen abhängigen Variablen können Sie benutzerdefinierte Scores für die einzelnen Kategorien der abhängigen Werte zuweisen. Die Scores definieren die Reihenfolge für die Kategorien der abhängigen Variablen und die Distanz zwischen diesen Kategorien. Mithilfe der Scores können Sie die relative Distanz zwischen ordinalen Werten vergrößern oder verkleinern sowie die Reihenfolge der Werte ändern.

### Für jede Kategorie ordinalen Rang verwenden

Die niedrigste Kategorie der abhängigen Variablen erhält den Score 1, die nächsthöhere Kategorie den Score 2 usw. Dies ist die Standardeinstellung.

### Benutzerdefiniert

Geben Sie je einen numerischen Score für jede im Raster aufgeführte Kategorie der abhängigen Variablen ein.

**Anmerkung:** In diesem Dialogfeld sind definierte Wertbeschriftungen für die abhängige Variable erforderlich. Das Dialogfeld ist erst dann verfügbar, wenn mindestens zwei Werte der kategorialen abhängigen Variablen eine Wertbeschriftung besitzen.

### Beispiel

*Tabelle 2. Angepasste Scorewerte*

| Wertbeschriftung  | Originalwert | Score |
|-------------------|--------------|-------|
| Ungelernt         | 1            | 1     |
| Gelernt/Werkstatt | 2            | 4     |
| Verwaltung        | 3            | 4,5   |
| Professional      | 4            | 7     |

Table 2. Angepasste Scorewerte (Forts.)

| Wertbeschriftung | Originalwert | Score |
|------------------|--------------|-------|
| Management       | 5            | 6     |

- Die Scores vergrößern die relative Distanz zwischen *Ungelernt* und *Gelernt/Werkstatt* und verringern die relative Distanz zwischen *Gelernt/Werkstatt* und *Verwaltung*.
- Die Scores kehren die Reihenfolge von *Management* und *Fachkraft* um.

### So legen Sie Scores fest

1. Wählen Sie im Hauptdialogfeld "Entscheidungsbaum" eine ordinale abhängige Variable mit mindestens zwei definierten Wertbeschriftungen aus.
2. Wählen Sie als **Aufbaumethode** die Option **CHAID** oder **Exhaustive CHAID** aus.
3. Klicken Sie auf **Scores**.

### Fehlende Werte

Im Dialogfeld "Fehlende Werte" steuern Sie die Behandlung benutzerdefiniert fehlender Werte für nominale unabhängige Variablen (Prädiktorvariablen).

- Benutzerdefiniert fehlende Werte für ordinale und metrische Variablen werden bei den verschiedenen Aufbaumethoden auf unterschiedliche Weise behandelt.
- Die Behandlung nominaler abhängiger Variablen wird im Dialogfeld "Kategorien" festgelegt. Weitere Informationen finden Sie in „Auswählen von Kategorien“ auf Seite 4.
- Bei ordinalen und metrischen abhängigen Variablen werden Fälle, bei denen systemdefiniert oder benutzerdefiniert fehlende Werte vorliegen, stets ausgeschlossen.

### Benutzerdefiniert fehlende Werte nominaler unabhängiger Variablen

#### Als fehlende Werte behandeln

Benutzerdefiniert fehlende Werte werden wie systemdefiniert fehlende Werte behandelt. Systemdefiniert fehlende Werte werden bei den verschiedenen Aufbaumethoden auf unterschiedliche Weise behandelt.

#### Als gültige Werte behandeln

Benutzerdefiniert fehlende Werte bei nominalen unabhängigen Variablen werden beim Aufbau und bei der Klassifizierung des Baums als normale Werte behandelt.

### Methodenabhängige Regeln

Einige (jedoch nicht alle) Werte für eine unabhängige Variable fehlen system- oder benutzerdefiniert:

- Bei CHAID und Exhaustive CHAID werden system- und benutzerdefiniert fehlende Werte für eine unabhängige Variable als eine einzige, kombinierte Kategorie in die Analyse aufgenommen. Bei metrischen und ordinalen unabhängigen Variablen werden mit den Algorithmen zunächst Kategorien mithilfe gültiger Werte erzeugt. Anschließend wird entschieden, ob die fehlende Kategorie mit der ähnlichsten (gültigen) Kategorie zusammengeführt oder als separate Kategorie beibehalten werden soll.
- Bei CRT und QUEST werden Fälle, bei denen Werte für eine unabhängige Variable fehlen, aus dem Vorgang des Baumaufbaus ausgeschlossen. Falls Surrogate in der Methode eingeschlossen sind, werden diese Fälle allerdings mithilfe von Surrogaten klassifiziert. Für nominale benutzerdefiniert fehlende Werte, die als fehlend behandelt werden, gilt dieselbe Vorgehensweise. Weitere Informationen finden Sie in „Surrogate“ auf Seite 9.

### So bestimmen Sie die Behandlung für nominale, unabhängige, benutzerdefiniert fehlende Werte:

1. Wählen Sie im Hauptdialogfeld "Entscheidungsbaum" mindestens eine nominale unabhängige Variable aus.
2. Klicken Sie auf **Fehlende Werte**.



## Speichern von Modellinformationen

Sie können die Informationen aus dem Modell als Variablen in der Arbeitsdatendatei speichern und auch das gesamte Modell im XML-Format (PMML) in einer externen Datei speichern.

### Gespeicherte Variablen

#### Endknotennummer

Endknoten, dem die einzelnen Fälle zugewiesen sind. Der Wert ist die Baumknotennummer.

#### Vorhergesagter Wert

Klasse (Gruppe) oder Wert für die abhängige Variable, der durch das Modell vorhergesagt wurde.

#### Vorhergesagte Wahrscheinlichkeiten

Wahrscheinlichkeit, die mit der Vorhersage des Modells verbunden ist. Für jede Kategorie der abhängigen Variablen wird je eine Variable gespeichert. Nicht verfügbar für metrische abhängige Variablen.

#### Stichprobenzuweisungen (Training/Tests)

Diese Variable zeigt bei der Split-Sample-Validierung, ob ein Fall in der Trainings- oder in der Teststichprobe verwendet wurde. Bei der Trainingsstichprobe ist der Wert gleich 1, bei der Teststichprobe dagegen gleich 0. Nur verfügbar, wenn die Split-Sample-Validierung ausgewählt ist. Weitere Informationen finden Sie in „Validierung“ auf Seite 5.

### Baummodell als XML exportieren

Sie können das gesamte Baummodell im XML-Format (PMML) speichern. Anhand dieser Modelldatei können Sie die Modellinformationen zu Scoring-Zwecken auf andere Datendateien anwenden.

#### Trainingsstichprobe

Schreibt das Modell in die angegebene Datei. Bei Bäumen mit Split-Sample-Validierung ist dies das Modell für die Trainingsstichprobe.

#### Teststichprobe

Schreibt das Modell für die Teststichprobe in die angegebene Datei. Nur verfügbar, wenn die Split-Sample-Validierung ausgewählt ist.

## Ausgabe

Die verfügbaren Ausgabeoptionen sind abhängig von der Aufbaumethode, dem Messniveau der abhängigen Variablen und anderen Einstellungen.

### Baumanzeige

Sie können das anfängliche Erscheinungsbild des Baums steuern oder auch die Baumanzeige ganz unterdrücken.

**Baum** Standardmäßig wird das Baumdiagramm in der Ausgabe eingeschlossen, die auf der Registerkarte "Ausgabe" angezeigt wird. Soll das Baumdiagramm nicht in der Ausgabe angezeigt werden, inaktivieren Sie diese Option.

#### Anzeige

Diese Optionen steuern das anfängliche Erscheinungsbild des Baumdiagramms auf der Registerkarte "Ausgabe". Diese Attribute können außerdem geändert werden, indem Sie den erzeugten Baum bearbeiten.

#### Ausrichtung

Der Baum kann wahlweise auf dem Kopf stehend (mit dem Stammknoten an oberster Stelle), von links nach rechts oder von rechts nach links angezeigt werden.

### **Knoteninhalt**

Die Knoten können Tabellen und/oder Diagramme enthalten. Bei kategorialen abhängigen Variablen zeigen die Tabellen die Häufigkeitszähler und die Prozentsätze; die Diagramme bestehen dabei aus Balkendiagrammen. Bei metrischen abhängigen Variablen zeigen die Tabellen die Mittelwerte, die Standardabweichungen, die Anzahl der Fälle und die vorhergesagten Werte. Die Diagramme bestehen dabei aus Histogrammen.

### **Skalierung**

Standardmäßig werden große Bäume so skaliert, dass der gesamte Baum auf der Seite dargestellt werden kann. Sie können eine benutzerdefinierte Skalierung bis 200 % angeben.

### **Statistik für unabhängige Variablen**

Bei CHAID und Exhaustive CHAID umfassen die Statistiken den *F*-Wert (metrische abhängige Variablen) bzw. den Chi-Quadrat-Wert (kategoriale abhängige Variablen), außerdem den Signifikanzwert und die Freiheitsgrade. Bei CRT wird der Verbesserungswert angezeigt. Bei QUEST werden der *F*-Wert, der Signifikanzwert und die Freiheitsgrade (für metrische und ordinale unabhängige Variablen) bzw. der Chi-Quadrat-Wert, der Signifikanzwert und die Freiheitsgrade (für nominale unabhängige Variablen) angezeigt.

### **Knotendefinitionen**

Die Knotendefinitionen zeigen den Wert oder die Werte der unabhängigen Variablen bei jeder Knotenaufteilung.

### **Baum im Tabellenformat**

Zusammenfassende Angaben für jeden Knoten im Baum: Nummer des übergeordneten Knotens, Statistik für unabhängige Variablen, Wert(e) der unabhängigen Variablen für den Knoten, Mittelwert und Standardabweichung für metrische abhängige Variablen bzw. Zählungen und Prozentsätze für kategoriale abhängige Variablen.

### **So steuern Sie die anfängliche Darstellung des Baums**

1. Klicken Sie im Hauptdialogfeld "Entscheidungsbaum" auf **Baum**.

### **Statistik**

Die verfügbaren Statistiktabelle sind abhängig vom Messniveau der abhängigen Variable, von der Aufbaumethode und anderen Einstellungen.

### **Modell**

#### **Zusammenfassung**

Die Zusammenfassung zeigt die verwendete Methode, die Variablen, die im Modell berücksichtigt sind, sowie die Variablen, die zwar angegeben, jedoch nicht in das Modell aufgenommen wurden.

**Risiko** Risikoschätzung und zugehöriger Standardfehler. Maß für die Vorhersagegenauigkeit des Baums.

- Bei kategorialen abhängigen Variablen ist die Risikoschätzung der Anteil der Fälle, die nach der Anpassung aufgrund der A-priori-Wahrscheinlichkeiten und Fehlklassifizierungskosten fehlerhaft klassifiziert wurden.
- Bei metrischen abhängigen Variablen ist die Risikoschätzung die Varianz innerhalb der Knoten.

#### **Klassifikationstabelle**

Bei kategorialen (nominalen, ordinalen) abhängigen Variablen zeigt diese Tabelle die Anzahl der Fälle in jeder Kategorie der abhängigen Variable, die korrekt bzw. fehlerhaft klassifiziert wurden. Nicht verfügbar für metrische abhängige Variablen.

### **Werte für Kosten, A-priori-Wahrscheinlichkeit, Score und Gewinn**

Bei kategorialen abhängigen Variablen zeigt diese Tabelle die Kostenwerte, die Werte für die A-priori-Wahrscheinlichkeiten, die Scores und die Profitwerte für die Analyse. Nicht verfügbar für metrische abhängige Variablen.

### **Unabhängige Variablen**

#### **Wichtigkeit für Modell**

Bei der CRT-Aufbaumethode wird jede unabhängige Variable (Prädiktor) gemäß ihrer Bedeutung für das Modell in eine Rangliste eingeordnet. Nicht verfügbar für QUEST- und CHAID-Methoden.

#### **Surrogate nach Aufteilung**

Bei den Aufbaumethoden CRT und QUEST werden die Surrogate für jede Aufteilung im Baum aufgeführt, sofern das Modell überhaupt Surrogate enthält. Nicht verfügbar für CHAID-Methoden. Weitere Informationen finden Sie in „Surrogate“ auf Seite 9.

### **Knotenleistung**

#### **Zusammenfassung**

Bei metrischen abhängigen Variablen enthält die Tabelle die Knotennummer, die Anzahl der Fälle und den Mittelwert für die abhängige Variable. Bei kategorialen abhängigen Variablen mit definierten Profiten zeigt die Tabelle die Knotennummer, die Anzahl der Fälle, den durchschnittlichen Profit sowie den Anlageertrag (ROI). Nicht verfügbar für kategoriale abhängige Variablen, bei denen keine Profite definiert sind. Weitere Informationen finden Sie in „Profite“ auf Seite 10.

#### **Nach Zielkategorie**

Bei kategorialen abhängigen Variablen mit definierten Zielkategorien enthält die Tabelle den prozentualen Gewinn, die Antworten in Prozent sowie den Indexprozentsatz (Anhebung) für die einzelnen Knoten- oder Perzentilgruppen. Für jede Zielkategorie wird eine separate Tabelle erstellt. Nicht verfügbar für metrische abhängige Variablen und kategoriale abhängige Variablen, bei denen jeweils keine Zielkategorien definiert sind. Weitere Informationen finden Sie in „Auswählen von Kategorien“ auf Seite 4.

**Zeilen** Die Tabellen mit der Knotenleistung können Ergebnisse nach Endknoten und/oder nach Perzentilen aufnehmen. Wenn Sie beide Elemente auswählen, werden je zwei Tabellen für jede Zielkategorie angelegt. Die Perzentiltabellen zeigen kumulative Werte für die einzelnen Perzentile auf der Grundlage der Sortierreihenfolge.

#### **Sortierreihenfolge**

Der Wert variiert je nach Messniveau der abhängigen Variablen und unterscheidet sich hinsichtlich der Gewinnzusammenfassung und der Gewinntabelle.

#### **Perzentilinkrement**

Bei Perzentiltabellen können Sie das Perzentilinkrement auswählen: 1, 2, 5, 10, 20 oder 25.

#### **Kumulative Statistik anzeigen**

Bei Endknotentabellen werden zusätzliche Spalten mit kumulativen Ergebnissen in die einzelnen Tabellen aufgenommen.

### **So wählen Sie die Statistikausgabe aus**

1. Klicken Sie im Hauptdialogfeld "Entscheidungsbaum" auf **Statistik**.

### **Diagramme**

Die verfügbaren Diagramme sind abhängig vom Messniveau der abhängigen Variable, von der Aufbaumethode und anderen Einstellungen.

#### **Wichtigkeit der unabhängigen Variablen im Modell**

Balkendiagramm über die Modellbedeutung nach unabhängiger Variable (Prädiktor). Nur für die CRT-Aufbaumethode verfügbar.

## Knotenleistung

### Gewinn

Der Gewinn ist der Prozentsatz aller Fälle in der Zielkategorie in jedem Knoten und wird wie folgt berechnet:  $(\text{Knotenziel } n / \text{Gesamtziel } n) \times 100$ . Das Gewinnendiagramm besteht aus einem Liniendiagramm kumulativer Perzentilgewinne, die wie folgt berechnet werden:  $(\text{Kumulatives Perzentilziel } n / \text{Gesamtziel } n) \times 100$ . Für jede Zielkategorie wird ein separates Liniendiagramm erstellt. Nur für kategoriale abhängige Variablen verfügbar, bei denen Zielkategorien definiert sind. Weitere Informationen finden Sie in „Auswählen von Kategorien“ auf Seite 4.

Das Gewinnendiagramm enthält dieselben Werte wie die Spalte *Gewinn (Prozent)* in der Tabelle "Gewinne für Perzentile"; hier werden ebenfalls kumulative Werte angezeigt.

**Index** Der Index ist das Verhältnis des Antwortprozentsatzes für die Zielkategorie im Knoten zum Gesamtantwortprozentsatz für die Zielkategorie der gesamten Stichprobe. Das Indexdiagramm ist ein Liniendiagramm kumulativer Perzentilindexwerte. Nur für kategoriale abhängige Variablen verfügbar. Der kumulative Perzentilindex wird wie folgt berechnet:  $(\text{Kumulative Perzentilantwort in Prozent} / \text{Gesamtantwort in Prozent}) \times 100$ . Für jede Zielkategorie wird ein separates Diagramm angelegt. Die Zielkategorien müssen definiert werden.

Das Indexdiagramm enthält dieselben Werte wie die Spalte *Index* in der Tabelle "Gewinne für Perzentile".

### Antwort

Der Prozentsatz der Fälle im Knoten, die der angegebenen Zielkategorie angehören. Das Antwortdiagramm besteht aus einem Liniendiagramm kumulativer Perzentilantworten, die wie folgt berechnet werden:  $(\text{Kumulatives Perzentilziel } n / \text{kumulative Perzentilgesamtzahl } n) \times 100$ . Nur für kategoriale abhängige Variablen verfügbar, bei denen Zielkategorien definiert sind.

Das Antwortdiagramm enthält dieselben Werte wie die Spalte *Antwort* in der Tabelle "Gewinne für Perzentile".

### Mittelwert

Liniendiagramm der kumulativen Perzentilmittelwerte für die abhängige Variable. Nur für metrische abhängige Variablen verfügbar.

### Durchschnittlicher Profit

Liniendiagramm des kumulativen durchschnittlichen Profits. Nur für kategoriale abhängige Variablen verfügbar, bei denen Profite definiert sind. Weitere Informationen finden Sie in „Profite“ auf Seite 10.

Das Diagramm für den durchschnittlichen Profit enthält dieselben Werte wie die Spalte *Profit* in der Tabelle "Gewinnzusammenfassung für Perzentile".

### Anlageertrag (ROI)

Liniendiagramm des kumulativen ROI (Anlageertrag). Der ROI wird als Verhältnis der Profite zu den Aufwendungen berechnet. Nur für kategoriale abhängige Variablen verfügbar, bei denen Profite definiert sind.

Das ROI-Diagramm enthält dieselben Werte wie die Spalte *ROI* in der Tabelle "Gewinnzusammenfassung für Perzentile".

### Perzentilinkrement

Bei allen Perzentildiagrammen steuert diese Einstellung die im Diagramm abgebildeten Perzentilinkremente: 1, 2, 5, 10, 20, or 25.

## So wählen Sie die Diagrammausgabe aus

1. Klicken Sie im Hauptdialogfeld "Entscheidungsbaum" auf **Diagramme**.

## Auswahl- und Scoring-Regeln

Im Dialogfeld "Regeln" legen Sie die Regeln für die Auswahl oder die Klassifizierung/Vorhersage mit der Befehlssyntax, als SQL-Anweisungen oder in natürlicher Sprache fest. Sie können diese Regeln auf der Registerkarte "Ausgabe" anzeigen und/oder in einer externen Datei speichern.

### Klassifizierungsregeln erstellen

Wählen Sie diese Option aus, um die Einstellung für die Auswahl und die Scoring-Regeln zu aktivieren.

### Syntax

Steuert die Form der Auswahlregeln sowohl für die Ausgabe auf der Registerkarte "Ausgabe" als auch beim Speichern in einer externen Datei.

### SPSS Statistics

Befehlssyntaxsprache. Die Regeln werden als Befehle ausgedrückt, die eine Filterbedingung zum Auswählen von Subsets mit Fällen definieren, oder auch als COMPUTE-Anweisungen, mit denen Fälle bewertet werden können.

**SQL** Um Datensätze auszuwählen oder aus einer Datenbank zu extrahieren oder um Werte für diese Datensätze zuzuweisen, werden Standard-SQL-Regeln erzeugt. Die erzeugten SQL-Regeln enthalten keine Tabellennamen oder andere Informationen zur Datenquelle.

**Text** Pseudocode in natürlicher Sprache. Regeln werden als Set logischer Wenn-dann-Anweisungen ausgedrückt, die die Klassifizierungen oder Vorhersagen des Modells für jeden Knoten beschreiben. Regeln in dieser Form können definierte Variablen- und Wertbeschriftungen oder auch Variablennamen und Datenwerte nutzen.

**Typ** Bei IBM® SPSS Statistics- und SQL-Regeln wird hiermit der Typ der erzeugten Regeln gesteuert: Auswahl- oder Scoringregeln.

### Fällen Werte zuweisen

Mit den Regeln können die Vorhersagen aus dem Modell zu Fällen zugewiesen werden, die die Kriterien für die Knotenzugehörigkeit erfüllen. Für jeden Knoten, der den Kriterien für die Knotenzugehörigkeit entspricht, wird eine separate Regel erzeugt.

### Fälle auswählen

Mit den Regeln können Fälle ausgewählt werden, die die Kriterien für die Knotenzugehörigkeit erfüllen. Bei IBM SPSS Statistics- und SQL-Regeln wird eine einzige Regel erzeugt, mit der alle Fälle ausgewählt werden, die den Auswahlkriterien entsprechen.

### Surrogate in SPSS Statistics- und SQL-Regeln einschließen

Bei CRT und QUEST können Sie Ersatzprädiktoren aus dem Modell in die Regeln aufnehmen. Regeln mit Surrogaten können recht komplex werden. Wenn Sie nur Informationen zu Konzepten Ihres Baums ableiten möchten, sollten Sie die Surrogate ausschließen. Wenn die Daten in den unabhängigen Variablen (Prädiktorvariablen) in bestimmten Fällen unvollständig sind und Regeln angelegt werden sollen, die den Baum getreu nachbilden, schließen Sie die Surrogate ein. Weitere Informationen finden Sie in „Surrogate“ auf Seite 9.

### Knoten

Steuert den Umfang der erzeugten Regeln. Für jeden Knoten im Umfang wird eine separate Regel erzeugt.

#### Alle Endknoten

Erzeugt Regeln für jeden Endknoten.

#### Beste Endknoten

Erzeugt Regeln für die besten  $n$  Endknoten auf der Grundlage der Indexwerte. Ist die Anzahl höher als die Anzahl der Endknoten im Baum, werden Regeln für alle Endknoten erzeugt.

**Beste Endknoten bis zu einem angegebenen Prozentsatz der Fälle.**

Erzeugt Regeln für Endknoten für die oberen  $n$  Prozent der Fälle auf der Grundlage der Indexwerte.

**Endknoten, deren Indexwert einen Trennwert erreicht oder überschreitet.**

Erzeugt Regeln für alle Endknoten, deren Indexwert größer-gleich dem angegebenen Wert ist. Ein Indexwert größer als 100 bedeutet, dass der Prozentsatz der Fälle in der Zielkategorie in diesem Knoten größer ist als der Prozentsatz im Stammknoten.

**Alle Knoten**

Erzeugt Regeln für alle Knoten.

**Hinweise:**

- Die Knotenauswahl auf der Grundlage der Indexwerte ist nur für kategoriale abhängige Variablen verfügbar, bei denen Zielkategorien definiert sind. Wenn Sie mehrere Zielkategorien angegeben haben, wird je ein Regelset für die einzelnen Zielkategorien erzeugt.
- Bei IBM SPSS Statistics- und SQL-Regeln zum Auswählen von Fällen (nicht bei Regeln zum Zuweisen von Werten) wird mit den Optionen **Alle Knoten** und **Alle Endknoten** eine Regel erzeugt, mit der alle Fälle in der Analyse ausgewählt werden.

**Regeln in Datei exportieren**

Speichert die Regeln in einer externen Textdatei.

Alternativ können Sie die Auswahl- und Scoring-Regeln interaktiv anhand ausgewählter Knoten im fertigen Baummodell erzeugen und speichern.

**Anmerkung:** Wenn Sie Regeln als Befehlssyntax auf eine andere Datendatei anwenden, müssen die Namen der Variablen in dieser Datendatei mit den Namen der unabhängigen Variablen im fertigen Modell identisch sein. Außerdem müssen die Variablen mit derselben Maßeinheit gemessen werden und dieselben benutzerdefiniert fehlenden Werte aufweisen (falls vorhanden).

**So legen Sie Auswahl- oder Scoring-Regeln fest**

1. Klicken Sie im Hauptdialogfeld "Entscheidungsbaum" auf **Regeln**.

---

## Bemerkungen

Die vorliegenden Informationen wurden für Produkte und Services entwickelt, die auf dem deutschen Markt angeboten werden. IBM stellt dieses Material möglicherweise auch in anderen Sprachen zur Verfügung. Für den Zugriff auf das Material in einer anderen Sprache kann eine Kopie des Produkts oder der Produktversion in der jeweiligen Sprache erforderlich sein.

Möglicherweise bietet IBM die in dieser Dokumentation beschriebenen Produkte, Services oder Funktionen in anderen Ländern nicht an. Informationen über die gegenwärtig im jeweiligen Land verfügbaren Produkte und Services sind beim zuständigen IBM Ansprechpartner erhältlich. Hinweise auf IBM Lizenzprogramme oder andere IBM Produkte bedeuten nicht, dass nur Programme, Produkte oder Services von IBM verwendet werden können. Anstelle der IBM Produkte, Programme oder Services können auch andere, ihnen äquivalente Produkte, Programme oder Services verwendet werden, solange diese keine gewerblichen oder anderen Schutzrechte von IBM verletzen. Die Verantwortung für den Betrieb von Produkten, Programmen und Services anderer Anbieter liegt beim Kunden.

Für in diesem Handbuch beschriebene Erzeugnisse und Verfahren kann es IBM Patente oder Patentanmeldungen geben. Mit der Auslieferung dieses Handbuchs ist keine Lizenzierung dieser Patente verbunden. Lizenzanforderungen sind schriftlich an folgende Adresse zu richten (Anfragen an diese Adresse müssen auf Englisch formuliert werden):

*IBM Director of Licensing  
IBM Europe, Middle East & Africa  
Tour Descartes  
2, avenue Gambetta  
92066 Paris La Defense  
France*

For license inquiries regarding double-byte (DBCS) information, contact the IBM Intellectual Property Department in your country or send inquiries, in writing, to:

INTERNATIONAL BUSINESS MACHINES CORPORATION 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 jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

Trotz sorgfältiger Bearbeitung können technische Ungenauigkeiten oder Druckfehler in dieser Veröffentlichung nicht ausgeschlossen werden. Die hier enthaltenen Informationen werden in regelmäßigen Zeitabständen aktualisiert und als Neuausgabe veröffentlicht. IBM kann ohne weitere Mitteilung jederzeit Verbesserungen und/oder Änderungen an den in dieser Veröffentlichung beschriebenen Produkten und/oder Programmen vornehmen.

Verweise in diesen Informationen auf Websites anderer Anbieter werden lediglich als Service für den Kunden bereitgestellt und stellen keinerlei Billigung des Inhalts dieser Websites dar. Das über diese Websites verfügbare Material ist nicht Bestandteil des Materials für dieses IBM Produkt. Die Verwendung dieser Websites geschieht auf eigene Verantwortung.

Werden an IBM Informationen eingesandt, können diese beliebig verwendet werden, ohne dass eine Verpflichtung gegenüber dem Einsender entsteht.

Lizenznehmer des Programms, die Informationen zu diesem Produkt wünschen mit der Zielsetzung: (i) den Austausch von Informationen zwischen unabhängig voneinander erstellten Programmen und ande-

ren Programmen (einschließlich des vorliegenden Programms) sowie (ii) die gemeinsame Nutzung der ausgetauschten Informationen zu ermöglichen, wenden sich an folgende Adresse:

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

Die Bereitstellung dieser Informationen kann unter Umständen von bestimmten Bedingungen - in einigen Fällen auch von der Zahlung einer Gebühr - abhängig sein.

Die Lieferung des in diesem Dokument beschriebenen Lizenzprogramms sowie des zugehörigen Lizenzmaterials erfolgt auf der Basis der IBM Rahmenvereinbarung bzw. der Allgemeinen Geschäftsbedingungen von IBM, der IBM Internationalen Nutzungsbedingungen für Programmpakete oder einer äquivalenten Vereinbarung.

Die angeführten Leistungsdaten und Kundenbeispiele dienen nur zur Illustration. Die tatsächlichen Ergebnisse beim Leistungsverhalten sind abhängig von der jeweiligen Konfiguration und den Betriebsbedingungen.

Alle Informationen zu Produkten anderer Anbieter stammen von den Anbietern der aufgeführten Produkte, deren veröffentlichten Ankündigungen oder anderen allgemein verfügbaren Quellen. IBM hat diese Produkte nicht getestet und kann daher keine Aussagen zu Leistung, Kompatibilität oder anderen Merkmalen machen. Fragen zu den Leistungsmerkmalen von Produkten anderer Anbieter sind an den jeweiligen Anbieter zu richten.

Aussagen über Pläne und Absichten von IBM unterliegen Änderungen oder können zurückgenommen werden und repräsentieren nur die Ziele von IBM.

Diese Veröffentlichung enthält Beispiele für Daten und Berichte des alltäglichen Geschäftsablaufs. Sie sollen nur die Funktionen des Lizenzprogramms illustrieren und können Namen von Personen, Firmen, Marken oder Produkten enthalten. Alle diese Namen sind frei erfunden; Ähnlichkeiten mit tatsächlichen Namen und Adressen sind rein zufällig.

#### COPYRIGHTLIZENZ:

Diese Veröffentlichung enthält Beispielanwendungsprogramme, die in Quellsprache geschrieben sind und Programmier Techniken in verschiedenen Betriebsumgebungen veranschaulichen. Sie dürfen diese Beispielprogramme kostenlos kopieren, ändern und verteilen, wenn dies zu dem Zweck geschieht, Anwendungsprogramme zu entwickeln, zu verwenden, zu vermarkten oder zu verteilen, die mit der Anwendungsprogrammierschnittstelle für die Betriebsumgebung konform sind, für die diese Beispielprogramme geschrieben werden. Diese Beispiele wurden nicht unter allen denkbaren Bedingungen getestet. Daher kann IBM die Zuverlässigkeit, Wartungsfreundlichkeit oder Funktion dieser Programme weder zusagen noch gewährleisten. Die Beispielprogramme werden ohne Wartung (auf "as-is"-Basis) und ohne jegliche Gewährleistung zur Verfügung gestellt. IBM übernimmt keine Haftung für Schäden, die durch die Verwendung der Beispielprogramme entstehen.

Kopien oder Teile der Beispielprogramme bzw. daraus abgeleiteter Code müssen folgenden Copyrightvermerk beinhalten:

© IBM 2019. Teile des vorliegenden Codes wurden aus Beispielprogrammen der IBM Corporation abgeleitet.

© Copyright IBM Corp. 1989 - 2019. Alle Rechte vorbehalten.



---

## Marken

IBM, das IBM Logo und ibm.com sind Marken oder eingetragene Marken der IBM Corp in den USA und/oder anderen Ländern. Weitere Produkt- und Servicennamen können Marken von IBM oder anderen Unternehmen sein. Eine aktuelle Liste der IBM Marken finden Sie auf der Webseite "Copyright and trademark information" unter [www.ibm.com/legal/copytrade.shtml](http://www.ibm.com/legal/copytrade.shtml).

Adobe, das Adobe-Logo, PostScript und das PostScript-Logo sind Marken oder eingetragene Marken der Adobe Systems Incorporated in den USA und/oder anderen Ländern.

Intel, das Intel-Logo, Intel Inside, das Intel Inside-Logo, Intel Centrino, das Intel Centrino-Logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium und Pentium sind Marken oder eingetragene Marken der Intel Corporation oder ihrer Tochtergesellschaften in den USA oder anderen Ländern.

Linux ist eine eingetragene Marke von Linus Torvalds in den USA und/oder anderen Ländern.

Microsoft, Windows, Windows NT und das Windows-Logo sind Marken der Microsoft Corporation in den USA und/oder anderen Ländern.

UNIX ist eine eingetragene Marke von The Open Group in den USA und anderen Ländern.

Java und alle auf Java basierenden Marken und Logos sind Marken oder eingetragene Marken der Oracle Corporation und/oder ihrer verbundenen Unternehmen.



---

# Index

## A

Ausblenden von Knoten  
im Unterschied zum Beschneiden 8

## B

Bäume 1  
A-priori-Wahrscheinlichkeit 10  
Anzahl der Ebenen einschränken 6  
Baumanzeige steuern 13  
Baumausrichtung 13  
Bauminhalt in einer Tabelle 13  
beschneiden 8  
CHAID-Aufbaukriterien 6  
CRT-Methode 7  
Diagramme 15  
Endknotenstatistik 14  
fehlende Werte 12  
Fehlklassifizierungskosten 9  
Fehlklassifizierungstabelle 14  
Indexwerte 14  
Intervalle für metrische unabhängige  
Variablen 7  
Knotengröße steuern 6  
Kreuzvalidierung 5  
Modellvariablen speichern 13  
Prädiktoreinfluss 14  
Profite 10  
Regeln erzeugen 17  
Risikoschätzung 14  
Scores 11  
Split-Sample-Validierung 5  
Verzweigungsstatistik ein- und aus-  
blenden 13  
Befehlssyntax  
Auswahl- und Scoring-Syntax für  
Klassifizierungsbäume erstellen 17

## C

CHAID 1  
Bonferroni-Korrektur 6  
erneut aufgeteilte, zusammengeführte  
Kategorien 6  
Intervalle für metrische unabhängige  
Variablen 7  
Kriterien für Aufteilen und Zusam-  
menführen 6  
Maximalzahl der Iterationen 6  
CRT 1  
beschneiden 8  
Unreinheitsmaße 7

## E

Entscheidungsbäume 1  
CHAID-Methode 1  
CRT-Methode 1

Entscheidungsbäume (*Forts.*)  
erste Variable in Modell aufnehmen  
lassen 1  
Exhaustive CHAID-Methode 1  
Messniveau 1  
QUEST-Methode 1, 8  
Entscheidungsbäume beschneiden  
im Unterschied zum Ausblenden von  
Knoten 8

## F

Fehlende Werte  
Bäume 12  
Fehlklassifizierung  
Bäume 14  
Kosten 9

## G

Gewichten von Fällen  
nicht ganzzahlige Gewichtungen in  
Entscheidungsbäumen 1  
Gini 7

## I

Indexwerte  
Bäume 14

## K

Knotennummer  
als Variable in Entscheidungsbäumen  
speichern 13  
Kosten  
Fehlklassifizierung 9  
Kreuzvalidierung  
Bäume 5

## M

Messniveau  
Entscheidungsbäume 1

## O

Ordinales Twoing 7

## P

Profite  
A-priori-Wahrscheinlichkeit 10  
Bäume 10, 14

## Q

QUEST 1, 8  
beschneiden 8

## R

Regeln  
Auswahl- und Scoring-Syntax für  
Klassifizierungsbäume erstellen 17  
Risikoschätzung  
Bäume 14

## S

Scores  
Bäume 11  
Signifikanzniveau für die Aufteilung von  
Knoten 8  
Split-Sample-Validierung  
Bäume 5  
SQL  
SQL-Syntax für Auswahl und Scoring  
erstellen 17  
Startwert für Zufallszahlen  
Entscheidungsbaumvalidierung 5  
Syntax  
Auswahl- und Scoring-Syntax für  
Klassifizierungsbäume erstellen 17

## T

Twoing 7

## U

Unreinheit  
CRT-Bäume 7

## V

Validierung  
Bäume 5  
Vorhergesagte Wahrscheinlichkeit  
als Variable in Entscheidungsbäumen  
speichern 13  
Vorhergesagte Werte  
als Variable in Entscheidungsbäumen  
speichern 13







Gedruckt in Deutschland