

IBM SPSS Modeler 14.2 – In-Database  
Mining-Handbuch



*Hinweis:* Lesen Sie vor der Verwendung dieser Informationen und des zugehörigen Produkts die allgemeinen Informationen unter Hinweise auf S. .

Dieses Dokument enthält eigentumsrechtlich geschützte Informationen von SPSS Inc., an IBM Company. Sie werden im Rahmen einer Lizenzvereinbarung bereitgestellt und sind durch Copyright-Gesetze geschützt. Die in dieser Publikation enthaltenen Informationen umfassen keinerlei Produktgewährleistungen und keine der Aussagen in diesem Handbuch darf als solche ausgelegt werden.

Wenn Sie Informationen an IBM bzw. SPSS senden, räumen Sie IBM und SPSS das nicht ausschließliche Recht ein, die Informationen in jeglicher Form zu verwenden bzw. weiterzugeben, die dem Unternehmen geeignet erscheint, ohne dass ihm daraus Verbindlichkeiten Ihnen gegenüber entstehen.

© **Copyright IBM Corporation 1994, 2011..**

---

# Vorwort

IBM® SPSS® Modeler ist die auf Unternehmensebene einsetzbare Data-Mining-Workbench von IBM Corp.. Mit SPSS Modeler können Unternehmen und Organisationen die Beziehungen zu ihren Kunden bzw. zu den Bürgern durch ein tief greifendes Verständnis der Daten verbessern. Organisationen benutzen die mithilfe von SPSS Modeler gewonnenen Erkenntnisse zur Bindung profitabler Kunden, zur Ermittlung von Cross-Selling-Möglichkeiten, zur Gewinnung neuer Kunden, zur Ermittlung von Betrugsfällen, zur Reduzierung von Risiken und zur Verbesserung der Verfügbarkeit öffentlicher Dienstleistungen.

Die visuelle Benutzeroberfläche von SPSS Modeler erleichtert die Anwendung des spezifischen Geschäftswissens der Benutzer, was zu leistungsstärkeren Vorhersagemodellen führt und die Zeit bis zur Lösungserstellung verkürzt. SPSS Modeler bietet zahlreiche Modellierungsverfahren, beispielsweise Algorithmen für Vorhersage, Klassifizierung, Segmentierung und Assoziationserkennung. Nach der Modellerstellung ermöglicht IBM® SPSS® Modeler Solution Publisher die unternehmensweite Bereitstellung für Entscheidungsträger oder in einer Datenbank.

## Über IBM Business Analytics

IBM Business Analytics-Software bietet vollständige, einheitliche und genaue Informationen, auf die Entscheidungsträger vertrauen, um die Unternehmensleistung zu steigern. Ein umfassendes Portfolio von Anwendungen für [Unternehmensinformationen](#), [Vorhersageanalysen](#), [Verwaltung der Finanzleistung und Strategie](#) sowie [Analysen](#) bietet sofort klare und umsetzbare Einblicke in die aktuelle Leistung und ermöglicht die Vorhersage zukünftiger Ergebnisse. In Kombination mit umfassenden Branchenlösungen, bewährten Vorgehensweisen und professionellen Dienstleistungen können Unternehmen jeder Größe optimale Produktivität erreichen, die Entscheidungsfindung zuverlässig automatisieren und bessere Ergebnisse erzielen.

Als Teil dieses Portfolios unterstützt die IBM SPSS Predictive Analytics-Software Unternehmen dabei, zukünftige Ereignisse vorherzusagen und aktiv auf diese Erkenntnisse zu reagieren, um bessere Geschäftsergebnisse zu erzielen. Kunden aus den Bereichen Wirtschaft, Behörden und Bildung aus aller Welt verlassen sich auf die IBM SPSS-Technologie. Sie bringt Ihnen beim Gewinnen, Halten und Ausbauen neuer Kundenbeziehungen einen Wettbewerbsvorteil und verringert gleichzeitig das Betrugs- sowie andere Risiken. Durch Integration der IBM SPSS-Software in den täglichen Betrieb können diese Unternehmen qualifizierte Vorhersagen treffen und dadurch die Entscheidungsfindung so ausrichten und automatisieren, dass Geschäftsziele erreicht werden und ein messbarer Wettbewerbsvorteil entsteht. Wenn Sie weitere Informationen wünschen oder einen Mitarbeiter kontaktieren möchten, ist dies unter <http://www.ibm.com/spss> möglich.

## Technischer Support

Kunden mit Wartungsvertrag können den technischen Support in Anspruch nehmen. Kunden können sich an den technischen Support wenden, wenn sie Hilfe bei der Arbeit mit IBM Corp.-Produkten oder bei der Installation in einer der unterstützten Hardware-Umgebungen benötigen. Die Kontaktdaten des Technischen Supports finden Sie auf der IBM Corp.-Website

unter <http://www.ibm.com/support>. Sie müssen bei der Kontaktaufnahme Ihren Namen, Ihre Organisation und Ihre Supportvereinbarung angeben.

---

# Inhalt

<b>1</b>	<b>Informationen zu IBM SPSS Modeler</b>	<b>1</b>
	IBM SPSS Modeler Server . . . . .	1
	IBM SPSS Modeler-Optionen . . . . .	2
	IBM SPSS Text Analytics . . . . .	2
	IBM SPSS Modeler-Dokumentation . . . . .	3
	Anwendungsbeispiele . . . . .	4
	Ordner "Demos" . . . . .	4
<b>2</b>	<b>In-Database Mining</b>	<b>6</b>
	Übersicht über die Datenbank-Modellierung. . . . .	6
	Was Sie brauchen . . . . .	7
	Modell konstruieren. . . . .	8
	Data Preparation (Vorbereitung von Daten) . . . . .	8
	Scores des Modells. . . . .	9
	Exportieren und Speichern von Datenbankmodellen . . . . .	10
	Modellkonsistenz. . . . .	11
	Anzeigen und Exportieren von generiertem SQL-Code . . . . .	11
<b>3</b>	<b>Datenbankmodellierung mit Microsoft Analysis Services</b>	<b>12</b>
	IBM SPSS Modeler und Microsoft Analysis Services . . . . .	12
	Anforderungen für die Integration mit Microsoft Analysis Services . . . . .	13
	Aktivieren der Integration mit Analysis Services . . . . .	15
	Erstellen von Modellen mit Analysis Services . . . . .	19
	Verwalten von Analysis Services-Modellen . . . . .	19
	Gemeinsame Einstellungen für alle Algorithmenknoten . . . . .	21
	MS Expertenoptionen für Entscheidungsbäume. . . . .	23
	MS Expertenoptionen für Clusterbildung . . . . .	24
	MS Expertenoptionen für Naive Bayes. . . . .	25
	MS Lineare Regression – Expertenoptionen. . . . .	26
	MS Neuronales Netzwerk – Expertenoptionen. . . . .	27
	MS Logistische Regression – Expertenoptionen. . . . .	28
	MS-Assoziationsregel-Knoten . . . . .	28
	MS Time Series-Knoten . . . . .	30
	MS-Sequenz-Clustering-Knoten. . . . .	34
	Scoring von Analysis Services-Modellen . . . . .	36
	Gemeinsame Einstellungen für alle Analysis Services-Modelle . . . . .	37

MS Zeitreihen-Modell-Nugget . . . . .	40
MS Sequenz-Clustering-Modell-Nugget . . . . .	44
Exportieren von Modellen und Generieren von Knoten . . . . .	44
Beispiele für das Mining mit Analysis Services . . . . .	44
Beispiel-Streams: Decision Trees (Entscheidungsbäume) . . . . .	45

## **4 Datenbank-Modellbildung mit Oracle Data Mining 53**

Informationen zu Oracle Data Mining . . . . .	53
Voraussetzungen für die Integration mit Oracle . . . . .	53
Aktivieren der Integration mit Oracle . . . . .	54
Modellbildung mit Oracle Data Mining . . . . .	57
Serveroptionen für Oracle-Modelle . . . . .	59
Fehlklassifizierungskosten . . . . .	60
Oracle Naive Bayes . . . . .	61
Optionen für Naive Bayes-Modelle . . . . .	61
Expertenoptionen für Naive Bayes . . . . .	62
Oracle Adaptive Bayes . . . . .	63
Optionen für Adaptive Bayes-Modelle . . . . .	64
Expertenoptionen für Adaptive Bayes . . . . .	65
Oracle Support Vector Machine (SVM) . . . . .	66
Optionen für Oracle SVM-Modelle . . . . .	66
Expertenoptionen für Oracle SVM . . . . .	68
Gewichtungsoptionen für Oracle SVM . . . . .	69
Verallgemeinerte lineare Modelle (GLM) von Oracle . . . . .	70
Optionen für Oracle GLM-Modelle . . . . .	71
Expertenoptionen für Oracle GLM . . . . .	72
Gewichtungsoptionen für Oracle GLM . . . . .	73
Oracle Decision Tree . . . . .	74
Optionen für Entscheidungsbaummodelle . . . . .	75
Expertenoptionen für Entscheidungsbäume . . . . .	76
Oracle O-Cluster . . . . .	77
Modelloptionen für O-Cluster . . . . .	77
Expertenoptionen für O-Cluster . . . . .	78
Oracle k-Means . . . . .	78
Optionen für das k-Means-Modell . . . . .	79
K-Means-Expertenoptionen . . . . .	80
Oracle Nonnegative Matrix Factorization (NMF) . . . . .	80
NMF-Modelloptionen . . . . .	81
NMF-Expertenoptionen . . . . .	82

Oracle Apriori . . . . .	83
Feldoptionen für A Priori . . . . .	83
Modelloptionen für A Priori . . . . .	86
Oracle Minimum Description Length (MDL). . . . .	87
MDL-Modelloptionen. . . . .	88
Oracle Attribute Importance (AI). . . . .	89
Alle Modelloptionen. . . . .	89
Alle Auswahloptionen . . . . .	90
AI-Modell-Nugget – Registerkarte “Modell” . . . . .	91
Verwalten von Oracle-Modellen . . . . .	92
Oracle-Modell-Nugget – Registerkarte “Server” . . . . .	92
Oracle-Modell-Nugget – Registerkarte “Übersicht”. . . . .	93
Oracle-Modell-Nugget – Registerkarte “Einstellungen” . . . . .	94
Auflisten der Oracle-Modelle. . . . .	94
Oracle Data Miner . . . . .	96
Vorbereitung der Daten . . . . .	98
Beispiele für Oracle Data Mining . . . . .	98
Beispiel-Stream: Hochladen von Daten . . . . .	99
Beispiel-Stream: Untersuchen von Daten . . . . .	100
Beispiel-Stream: Erstellen des Modells . . . . .	101
Beispiel-Stream: Evaluieren des Modells. . . . .	102
Beispiel-Stream: Bereitstellen des Modells . . . . .	105

## **5 Datenbankmodellierung mit IBM InfoSphere Warehouse 106**

IBM InfoSphere Warehouse und IBM SPSS Modeler . . . . .	106
Anforderungen für die Integration mit InfoSphere Warehouse . . . . .	106
Aktivieren der Integration mit InfoSphere Warehouse . . . . .	107
Modellbildung mit InfoSphere Warehouse Data Mining . . . . .	115
Modell-Scoring und -Deployment . . . . .	115
Verwalten von DB2-Modellen . . . . .	117
Auflistung der Datenbankmodelle . . . . .	117
Durchsuchen von Modellen. . . . .	118
Exportieren von Modellen und Generieren von Knoten. . . . .	119
Knoteneinstellungen, die für alle Algorithmen gelten . . . . .	119
ISW-Entscheidungsbaum . . . . .	123
Optionen für ISW-Entscheidungsbaummodelle . . . . .	124
Expertenoptionen für ISW-Entscheidungsbäume . . . . .	125
ISW-Assoziation . . . . .	125
Optionen für ISW-Assoziationsmodelle . . . . .	126
Expertenoptionen für ISW-Assoziationen. . . . .	127

ISW-Taxonomieoptionen . . . . .	128
ISW-Sequenz . . . . .	131
Optionen für ISW-Sequenzmodelle . . . . .	132
Expertenoptionen für ISW-Sequenzen . . . . .	133
ISW-Regression . . . . .	134
Optionen für ISW-Regreessionsmodelle. . . . .	136
Expertenoptionen für ISW-Regressionen . . . . .	137
ISW Clustering . . . . .	139
Modelloptionen für ISW Clustering . . . . .	140
Expertenoptionen für ISW Clustering . . . . .	141
ISW Naive Bayes . . . . .	144
Optionen für ISW Naive Bayes-Modelle. . . . .	144
ISW Logistische Regression . . . . .	145
Optionen für logistische ISW-Regreessionsmodelle. . . . .	145
ISW Time Series . . . . .	146
ISW-Zeitreihen – Feldoptionen. . . . .	147
ISW-Zeitreihenmodelle – Optionen . . . . .	148
ISW-Zeitreihen – Expertenoptionen . . . . .	149
Anzeigen von ISW-Zeitreihenmodellen . . . . .	150
ISW Data Mining Modell-Nuggets . . . . .	151
ISW-Modell-Nugget – Registerkarte “Server” . . . . .	151
ISW-Modell-Nugget – Registerkarte “Einstellungen” . . . . .	152
ISW-Modell-Nugget – Registerkarte “Übersicht” . . . . .	153
Beispiele für ISW Data Mining . . . . .	154
Beispiel-Stream: Hochladen von Daten . . . . .	154
Beispiel-Stream: Untersuchen von Daten . . . . .	155
Beispiel-Stream: Erstellen des Modells . . . . .	156
Beispiel-Stream: Evaluieren des Modells. . . . .	157
Beispiel-Stream: Bereitstellen des Modells . . . . .	159

## **6 Datenbankmodellierung mit IBM Netezza Analytics 161**

IBM SPSS Modeler und IBM Netezza Analytics . . . . .	161
Voraussetzungen für die Integration mit IBM Netezza Analytics . . . . .	161
Aktivieren der Integration mit IBM Netezza Analytics . . . . .	162
Erstellen von Modellen mit IBM Netezza Analytics . . . . .	164
Serveroptionen für Netezza-Modelle . . . . .	164
Modelloptionen für Netezza-Modelle . . . . .	166
Netezza-Entscheidungsbäume . . . . .	167
Instanzgewichtungen und Klassengewichtungen. . . . .	168

Optionen für Netezza-Entscheidungsbaumfelder . . . . .	168
Erstellungsoptionen für Netezza-Entscheidungsbäume . . . . .	170
Netezza-K-Means. . . . .	175
K-Means-Feldoptionen von Netezza. . . . .	175
K-Means-Erstellungsoptionen von Netezza . . . . .	177
Verwalten von IBM Netezza Analytics-Modellen. . . . .	178
Entscheidungsbaummodell-Nuggets von Netezza . . . . .	178
Netezza-Modell-Nugget vom Typ "K-Means" . . . . .	181

## ***Anhang***

<b><i>A Hinweise</i></b>	<b>185</b>
--------------------------	------------

<b><i>Index</i></b>	<b>188</b>
---------------------	------------



# **Informationen zu IBM SPSS Modeler**

IBM® SPSS® Modeler ist ein Set von Data-Mining-Tools, mit dem Sie auf der Grundlage Ihres Geschäftswissens schnell und einfach Vorhersagemodelle erstellen und zur Erleichterung der Entscheidungsfindung in die Betriebsabläufe einbinden können. SPSS Modeler, das auf der Grundlage des den Industrienormen entsprechenden Modells CRISP-DM entwickelt wurde, unterstützt den gesamten Data-Mining-Prozess, von den Daten bis hin zu besseren Geschäftsergebnissen.

SPSS Modeler bietet eine Vielzahl von Modellbildungsmethoden, die aus dem maschinellen Lernen, der künstlichen Intelligenz und der Statistik stammen. Mit den in der Modellierungspalette verfügbaren Methoden können Sie aus Ihren Daten neue Informationen ableiten und Vorhersagemodelle erstellen. Jede Methode besitzt ihre Stärken und eignet sich besonders für bestimmte Problemtypen.

SPSS Modeler kann als Standalone-Produkt oder in Verbindung mit SPSS Modeler Server erworben werden. Außerdem ist eine Reihe von Zusatzoptionen verfügbar, die in den folgenden Abschnitten kurz dargelegt werden. Weitere Informationen finden Sie unter <http://www.ibm.com/software/analytics/spss/products/modeler/>.

## **IBM SPSS Modeler Server**

SPSS Modeler verwendet eine Client/Server-Architektur zur Verteilung von Anforderungen für ressourcenintensive Vorgänge an leistungsstarke Serversoftware, wodurch bei größeren Daten-Sets eine schnellere Leistung erzielt werden kann. Neben den hier aufgeführten Produkten können auch weitere Produkte bzw. Aktualisierungen verfügbar sein. Weitere Informationen finden Sie unter <http://www.ibm.com/software/analytics/spss/products/modeler/>.

**SPSS Modeler.** SPSS Modeler Clementine Client ist eine im Funktionsumfang vollständige Version des installierten Produkts und kann auf dem Desktop-Computer des Benutzers ausgeführt werden. Es kann im lokalen Modus als Standalone-Produkt oder im verteilten Modus zusammen mit IBM® SPSS® Modeler Server verwendet werden, um im Falle von großen Daten-Sets die Leistung zu verbessern.

**SPSS Modeler Server.** SPSS Modeler Server wird ständig im verteilten Analysemodus zusammen mit einer oder mehreren IBM® SPSS® Modeler-Installationen ausgeführt, wodurch eine herausragende Leistung bei großen Daten-Sets erzielt werden kann, da speicherintensive Operationen auf dem Server durchgeführt werden können, ohne Daten auf den Client-Computer herunterzuladen. SPSS Modeler Server bietet außerdem Unterstützung für SQL-Optimierung sowie Funktionen zur Modellierung innerhalb der Datenbank, wodurch Leistungsfähigkeit und Automatisierung weiter verbessert werden. Es muss mindestens eine SPSS Modeler-Installation vorhanden sein, um eine Analyse durchzuführen.

## **IBM SPSS Modeler-Optionen**

Die folgenden Komponenten und Funktionen können separat erworben und für die Verwendung mit SPSS Modeler lizenziert werden. Beachten Sie, dass zu einem späteren Zeitpunkt möglicherweise noch weitere Produkte und Updates erhältlich sind. Weitere Informationen finden Sie unter <http://www.ibm.com/software/analytics/spss/products/modeler/>.

- SPSS Modeler Server-Zugriff, der eine bessere Skalierbarkeit und bessere Leistungsfähigkeit bei großen Daten-Sets bietet sowie Unterstützung für SQL-Optimierung und Funktionen zur Modellierung innerhalb der Datenbank.
- SPSS Modeler Solution Publisher, für Scoring in Echtzeit oder automatisiertes Scoring außerhalb der SPSS Modeler-Umgebung. Für weitere Informationen siehe Thema **IBM SPSS Modeler Solution Publisher** in Kapitel 2 in *IBM SPSS Modeler 14.2 Solution Publisher*.
- Adapter zum Deployment für IBM SPSS Collaboration and Deployment Services oder die Thin-Client-Anwendung IBM SPSS Modeler Advantage. Für weitere Informationen siehe Thema **Speichern und Bereitstellen von IBM SPSS Collaboration and Deployment Services Repository-Objekten** in Kapitel 9 in *IBM SPSS Modeler 14.2- Benutzerhandbuch*.

## **IBM SPSS Text Analytics**

IBM® SPSS® Text Analytics ist ein vollständig integriertes Zusatzprodukt für SPSS Modeler, das hoch entwickelte linguistische Technologien und die Verarbeitung natürlicher Sprache (Natural Language Processing, NLP) benutzt, um eine schnelle Verarbeitung einer großen Vielfalt an unstrukturierten Textdaten zu ermöglichen, um die Schlüsselkonzepte zu extrahieren und zu ordnen und um diese Konzepte in Kategorien zusammenzufassen. Extrahierte Konzepte und Kategorien können mit bestehenden strukturierten Daten, beispielsweise demografischen Informationen, kombiniert und mithilfe der vollständigen Suite der Data-Mining-Tools von IBM® SPSS® Modeler auf die Modellierung angewendet werden, um bessere und fokussiertere Entscheidungen zu ermöglichen.

- Der Text-Mining-Knoten bietet die Modellierung von Konzepten und Kategorien sowie eine interaktive Workbench, in der Sie eine erweiterte Untersuchung von Textlinks und Clustern durchführen, Ihre eigenen Kategorien erstellen und die Vorlagen für linguistische Ressourcen verfeinern können.
- Eine Reihe von Importformaten wird unterstützt, darunter Blogs und andere webbasierte Quellen.
- Benutzerdefinierte Vorlagen, Bibliotheken und Wörterbücher für bestimmte Domänen, wie CRM und Genomforschung, sind ebenfalls eingeschlossen.

*Anmerkung:* Für den Zugriff auf diese Komponente ist eine separate Lizenz erforderlich. Weitere Informationen finden Sie unter <http://www.ibm.com/software/analytics/spss/products/modeler/>.

## **IBM SPSS Modeler-Dokumentation**

Die vollständige Dokumentation im Online-Hilfe-Format finden Sie im Hilfe-Menü von SPSS Modeler. Dazu gehören die Dokumentation für SPSS Modeler, SPSS Modeler Server und SPSS Modeler Solution Publisher sowie das Anwendungshandbuch und weiteres Material zur Unterstützung.

Die vollständige Dokumentation für die einzelnen Produkte im PDF-Format finden Sie im Ordner *Documentation* auf der jeweiligen Produkt-DVD.

- **IBM SPSS Modeler-Benutzerhandbuch.** Allgemeine Einführung in die Verwendung von SPSS Modeler, in der u. a. die Erstellung von Daten-Streams, der Umgang mit fehlenden Werten, die Erstellung von CLEM-Ausdrücken, die Arbeit mit Projekten und Berichten sowie das Packen von Streams für das Deployment in IBM SPSS Collaboration and Deployment Services, Predictive Applications (Prognoseanwendungen) oder IBM SPSS Modeler Advantage beschrieben werden.
- **Quellen-, Prozess- und Ausgabeknoten in IBM SPSS Modeler.** Beschreibung aller Knoten, die zum Lesen, zum Verarbeiten und zur Ausgabe von Daten in verschiedenen Formaten verwendet werden. Im Grunde sind sie alle Knoten, mit Ausnahme der Modellierungsknoten.
- **IBM SPSS Modeler Modellierungsknoten.** Beschreibungen sämtlicher für die Erstellung von Data Mining-Modellen verwendeter Knoten. IBM® SPSS® Modeler bietet eine Vielzahl von Modellbildungsmethoden, die aus dem maschinellen Lernen, der künstlichen Intelligenz und der Statistik stammen. [Für weitere Informationen siehe Thema Überblick über Modellierungsknoten in Kapitel 3 in IBM SPSS Modeler 14.2-Modellierungsknoten.](#)
- **IBM SPSS Modeler-Algorithmushandbuch.** Beschreibung der mathematischen Grundlagen der in SPSS Modeler verwendeten Modellierungsmethoden.
- **IBM SPSS Modeler-Anwendungshandbuch.** Die Beispiele in diesem Handbuch bieten eine kurze, gezielte Einführung in bestimmte Modellierungsmethoden und -verfahren. Eine Online-Version dieses Handbuchs kann auch über das Hilfe-Menü aufgerufen werden. [Für weitere Informationen siehe Thema Anwendungsbeispiele in IBM SPSS Modeler 14.2-Benutzerhandbuch.](#)
- **Skripterstellung und Automatisierung in IBM SPSS Modeler.** Informationen zur Automatisierung des Systems über Skripterstellung, einschließlich der Eigenschaften, die zur Bearbeitung von Knoten und Streams verwendet werden können.
- **IBM SPSS Modeler Deployment-Handbuch.** Informationen zum Ausführen von SPSS Modeler-Streams und -Szenarien als Schritte bei der Verarbeitung von Jobs im IBM® SPSS® Collaboration and Deployment Services Deployment Manager.
- **IBM SPSS Modeler CLEF-Entwicklerhandbuch.** CLEF bietet die Möglichkeit, Drittanbieterprogramme, wie Datenverarbeitungsroutinen oder Modellierungsalgorithmen, als Knoten in SPSS Modeler zu integrieren.
- **In-Database Mining-Handbuch für IBM SPSS Modeler.** Informationen darüber, wie Sie Ihre Datenbank dazu einsetzen, die Leistung zu verbessern, und wie Sie die Palette der Analysefunktionen über Drittanbieteralgorithmen erweitern.
- **IBM SPSS Modeler Server- und -Leistungshandbuch.** Informationen zur Konfiguration und Verwaltung von IBM® SPSS® Modeler Server.

- **IBM SPSS Modeler Administration Console – Benutzerhandbuch.** Informationen zur Installation und Nutzung der Konsolen-Benutzeroberfläche zur Überwachung und Konfiguration von SPSS Modeler Server. Die Konsole ist als Plugin für die Deployment Manager-Anwendung implementiert.
- **IBM SPSS Modeler Solution Publisher-Handbuch.** SPSS Modeler Solution Publisher ist eine Zusatzkomponente, mit der Unternehmen Streams zur Verwendung außerhalb der SPSS Modeler-Standardumgebung veröffentlichen können.
- **IBM SPSS Modeler-Handbuch zu CRISP-DM.** Schritt-für-Schritt-Anleitung für das Data-Mining mit SPSS Modeler unter Verwendung der CRISP-DM-Methode.

## **Anwendungsbeispiele**

Mit den Data-Mining-Tools in SPSS Modeler kann eine große Bandbreite an geschäfts- und unternehmensbezogenen Problemen gelöst werden; die Anwendungsbeispiele dagegen bieten jeweils eine kurze, gezielte Einführung in spezielle Modellierungsmethoden und -verfahren. Die hier verwendeten Daten-Sets sind viel kleiner als die riesigen Datenbestände, die von einigen Data-Mining-Experten verwaltet werden müssen, die zugrunde liegenden Konzepte und Methoden sollten sich jedoch auch auf reale Anwendungen übertragen lassen.

Sie können auf die Beispiele zugreifen, indem Sie im Menü “Hilfe” in SPSS Modeler auf die Option Anwendungsbeispiele klicken. Die Datendateien und Beispiel-Streams wurden im Ordner *Demos*, einem Unterordner des Produktinstallationsverzeichnis, installiert. [Für weitere Informationen siehe Thema Ordner “Demos” in IBM SPSS Modeler 14.2- Benutzerhandbuch.](#)

**Beispiele für die Datenbank-Modellierung.** Die Beispiele finden Sie im *IBM SPSS Modeler In-Database Mining-Handbuch*.

**Skriptbeispiele.** Die Beispiele finden Sie im *IBM SPSS Modeler Handbuch für die Skripterstellung und Automatisierung*.

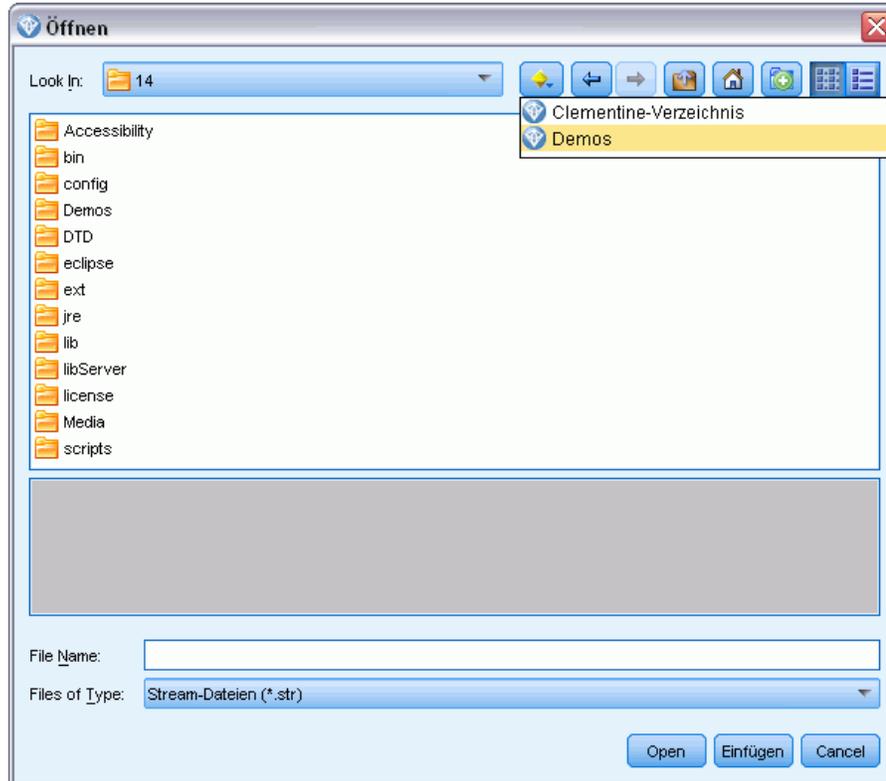
## **Ordner “Demos”**

Die in den Anwendungsbeispielen verwendeten Datendateien und Beispiel-Streams wurden im Ordner *Demos*, einem Unterordner des Produktinstallationsverzeichnis, installiert. Auf diesen Ordner können Sie auch über die Programmgruppe IBM SPSS Modeler 14.2 im

Windows-Startmenü oder durch Klicken auf *Demos* in der Liste der zuletzt angezeigten Verzeichnisse im Dialogfeld "Datei öffnen" zugreifen.

**Abbildung 1-1**

*Auswahl des Ordners "Demos" in der Liste der zuletzt angezeigten Verzeichnisse*



# ***In-Database Mining***

## ***Übersicht über die Datenbank-Modellierung***

IBM® SPSS® Modeler Server unterstützt die Integration mit Data-Mining-Tools und Daten-Modellierungstools von Datenbankherstellern wie Oracle Data Miner, IBM DB2 InfoSphere Warehouse und Microsoft Analysis Services. Sie können Modelle in der Datenbank erstellen, bewerten und speichern – ohne dazu die IBM® SPSS® Modeler-Anwendung verlassen zu müssen. Damit können Sie die analytischen Funktionen und die Benutzerfreundlichkeit des SPSS Modeler-Desktops mit der Leistungsstärke einer Datenbank kombinieren und gleichzeitig die datenbankeigenen Algorithmen nutzen, die von diesen Herstellern angeboten werden. Die Modelle werden innerhalb der Datenbank erstellt. Anschließend können Sie sie auf normale Weise über die SPSS Modeler-Benutzeroberfläche durchsuchen und scoren und bei Bedarf ihr Deployment mithilfe von IBM® SPSS® Modeler Solution Publisher durchführen. Die unterstützten Algorithmen befinden sich in der Datenbank-Modellierungspalette von SPSS Modeler.

Der Zugriff auf datenbankeigene Algorithmen mithilfe von SPSS Modeler bietet mehrere Vorteile:

- Datenbankeigene Algorithmen sind häufig eng mit dem Datenbankserver integriert und bieten u. U. eine verbesserte Leistung.
- Modelle, die in der Datenbank erstellt und gespeichert werden, können einfacher verwendet und für alle Anwendungen, die Zugriff auf die Datenbank haben, freigegeben werden.

**SQL-Erzeugung.** Die Modellierung innerhalb der Datenbank ist nicht dasselbe wie die SQL-Erzeugung, die auch als SQL-Optimierung bzw. "SQL-Pushback" bekannt ist. Mit dieser Funktion können Sie SQL-Anweisungen für systemeigene SPSS Modeler-Operationen erstellen, die dann zur Leistungsverbesserung per Pushback in die Datenbank zurückübertragen werden können. Die Merge-, Aggregat- und Auswahlknoten generieren beispielsweise jeweils SQL-Code, der auf diese Weise per Pushback an die Datenbank zurückübertragen werden kann. Die Verwendung von SQL-Erzeugung in Verbindung mit Datenbankmodellierung kann zu Streams führen, die von Anfang bis Ende in der Datenbank ausgeführt werden können, was erhebliche Leistungssteigerungen gegenüber in SPSS Modeler ausgeführten Streams mit sich bringt. [Für weitere Informationen siehe Thema SQL-Optimierung in Kapitel 6 in IBM SPSS Modeler Server 14.2-Verwaltungs- und Leistungshandbuch.](#)

*Hinweis:* Für Datenbankmodellierung und SQL-Optimierung muss auf dem IBM® SPSS® Modeler-Computer die SPSS Modeler Server-Konnektivität aktiviert sein. Wenn diese Einstellung aktiviert ist, können Sie auf Datenbankalgorithmen zugreifen, SQL direkt aus SPSS Modeler per Pushback übertragen und auf SPSS Modeler Server zugreifen. Um den aktuellen Lizenzstatus zu überprüfen, klicken Sie im Menü "Hilfe" in SPSS Modeler auf Info. [Für weitere Informationen siehe Thema Verbindung mit IBM SPSS Modeler Server in Kapitel 3 in IBM SPSS Modeler 14.2- Benutzerhandbuch.](#)



- In IBM® SPSS® Modeler muss im Dialogfeld “Hilfsprogramme” die Datenbank-Modellierung aktiviert sein (Extras > Hilfsprogramme).
- In IBM® SPSS® Modeler sowie in IBM® SPSS® Modeler Server (sofern verwendet) sollten im Dialogfeld “Benutzeroptionen” die Einstellungen SQL generieren und SQL-Optimierung aktiviert sein. [Für weitere Informationen siehe Thema Leistung und Optimierung in Kapitel 4 in IBM SPSS Modeler Server 14.2-Verwaltungs- und Leistungshandbuch.](#) Beachten Sie, dass SQL-Optimierung für die Datenbank-Modellierung nicht zwingend erforderlich ist, jedoch aus Leistungsgründen dringend empfohlen wird.

*Hinweis:* Für Datenbankmodellierung und SQL-Optimierung muss auf dem SPSS Modeler-Computer die SPSS Modeler Server-Konnektivität aktiviert sein. Wenn diese Einstellung aktiviert ist, können Sie auf Datenbankalgorithmen zugreifen, SQL direkt aus SPSS Modeler per Pushback übertragen und auf SPSS Modeler Server zugreifen. Um den aktuellen Lizenzstatus zu überprüfen, klicken Sie im Menü “Hilfe” in SPSS Modeler auf Info. [Für weitere Informationen siehe Thema Verbindung mit IBM SPSS Modeler Server in Kapitel 3 in IBM SPSS Modeler 14.2- Benutzerhandbuch.](#)

Detaillierte Informationen finden Sie in den nachfolgenden, herstellerspezifischen Abschnitten.

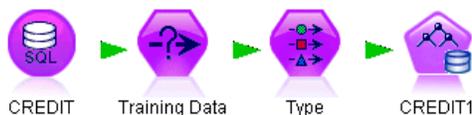
## Modell konstruieren

Der Vorgang des Erstellens und Scorens von Modellen mithilfe von Datenbankalgorithmen ähnelt anderen Data Mining-Typen in IBM® SPSS® Modeler. Die allgemeinen Abläufe bei der Arbeit mit Knoten und Modellierungs-”Nuggets” ähneln der Arbeit mit anderen Streams in SPSS Modeler. Der einzige Unterschied liegt darin, dass die eigentliche Verarbeitung und Modellerstellung an die Datenbank zurückgegeben werden.

Der folgende Stream beispielsweise ist konzeptuell mit anderen Daten-Streams in SPSS Modeler identisch. Dieser Stream führt jedoch alle Vorgänge in einer Datenbank aus, auch die Modellerstellung mit dem Knoten für Microsoft Decision Trees. Wenn Sie den Stream ausführen, weist SPSS Modeler die Datenbank an, das aus dem Stream resultierende Modell zu erstellen und zu speichern und die zugehörigen Details nach SPSS Modeler herunterzuladen.

Abbildung 2-3

*Beispiel-Stream für die Datenbank-Modellierung. Die violett eingezeichneten Knoten werden in der Datenbank ausgeführt.*



## Data Preparation (Vorbereitung von Daten)

Unabhängig davon, ob datenbankinterne Algorithmen verwendet werden oder nicht, sollten Sie die Datenvorbereitungsaufgaben nach Möglichkeit an die Datenbank zurückgeben, um so die Leistung zu steigern.

- Sind die Originaldaten in der Datenbank gespeichert, besteht das Ziel darin, die Daten dort zu behalten. Stellen Sie hierzu sicher, dass alle erforderlichen aufwärts liegenden Operationen in SQL konvertiert werden können. Auf diese Weise vermeiden Sie, dass die Daten nach IBM®

SPSS® Modeler— heruntergeladen werden (und somit ein Engpass entsteht, der den gesamten Leistungszuwachs zunichte machen würde), und Sie sorgen dafür, dass der gesamte Stream in der Datenbank ausgeführt wird. [Für weitere Informationen siehe Thema SQL-Optimierung in Kapitel 6 in IBM SPSS Modeler Server 14.2-Verwaltungs- und Leistungshandbuch.](#)

- Sind die Originaldaten *nicht* in der Datenbank gespeichert, kann die Datenbank-Modellierung dennoch verwendet werden. In diesem Fall wird die Datenvorbereitung in SPSS Modeler ausgeführt und die vorbereiteten Daten werden automatisch an die Datenbank zum Erstellen des Modells hochgeladen.

## Scoren des Modells

Modelle, die in IBM® SPSS® Modeler mit In-Database Mining generiert werden, unterscheiden sich von regulären SPSS Modeler-Modellen. Obwohl sie im Model Manager als generierte Modell-”Nuggets” angezeigt werden, handelt es sich jedoch tatsächlich um Remote-Modelle, die auf dem entfernten Data-Mining- oder Datenbankserver gespeichert werden. In SPSS Modeler werden lediglich Verweise auf die Remote-Modelle angezeigt. Mit anderen Worten: Das Modell, das Sie in SPSS Modeler sehen, ist eine “Modellhülle”, die Informationen wie den Hostnamen des Datenbankservers, den Datenbanknamen sowie den Modellnamen enthält. Dies ist eine wichtige Unterscheidung, die beim Durchsuchen und Scoren von Modellen, die mit datenbankinternen Algorithmen erstellt wurden, berücksichtigt werden muss.

Abbildung 2-4

Generiertes Modell-”Nugget” für Microsoft Decision Trees



Sobald Sie ein Modell erstellt haben, können Sie es wie jedes andere in SPSS Modeler generierte Modell dem Stream zum Zwecke des Scoring hinzufügen. Das gesamte Scoring erfolgt innerhalb der Datenbank, auch wenn dies bei weiter oben im Stream liegenden Operationen nicht der Fall ist. (Weiter oben im Stream liegende Operationen können weiterhin nach Möglichkeit per Pushback an die Datenbank zurückübertragen werden, um die Leistungsfähigkeit zu verbessern, dies ist jedoch keine Voraussetzung, damit das Scoring stattfinden kann.) Außerdem können Sie das generierte Modell in den meisten Fällen mithilfe des vom Datenbankanbieter bereitgestellten Standard-Browsers durchsuchen.

Für das Durchsuchen und das Scoring ist jeweils eine Live-Verbindung zu dem Server, auf dem Oracle Data Miner, IBM DB2 InfoSphere Warehouse bzw. Microsoft Analysis Services ausgeführt wird, erforderlich.

### Anzeigen von Ergebnissen und Festlegen von Einstellungen

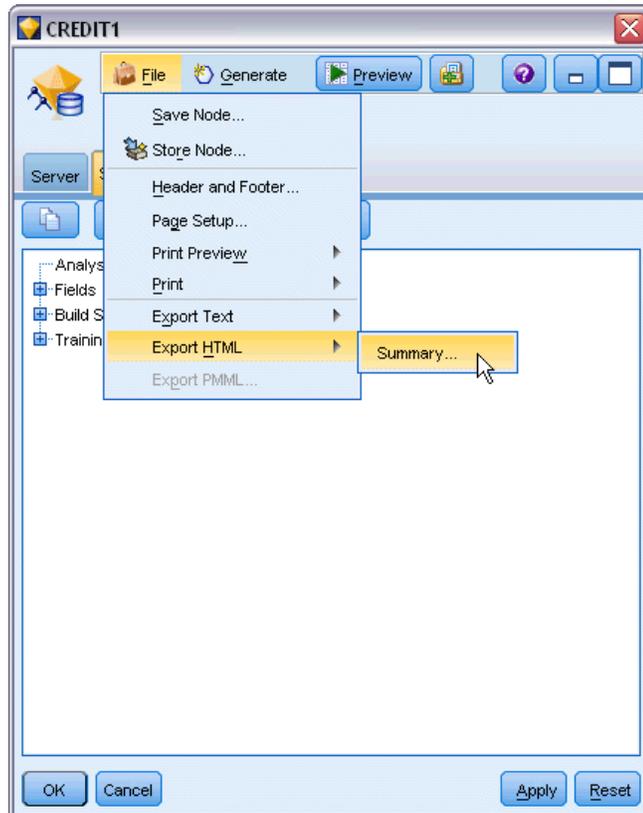
Doppelklicken Sie zum Anzeigen von Ergebnissen und zum Festlegen von Einstellungen für das Scoring auf das Modell im Stream-Zeichenbereich. Sie können auch mit der rechten Maustaste auf das Modell klicken und Durchsuchen oder Bearbeiten wählen. Bestimmte Einstellungen hängen vom Typ des Modells ab.

## Exportieren und Speichern von Datenbankmodellen

Datenbankmodelle können wie andere in IBM® SPSS® Modeler erstellte Modelle und Übersichten mit den Optionen im Menü “Datei” aus dem Modellbrowser exportiert werden.

Abbildung 2-5

Exportieren einer Microsoft Decision Trees-Modellübersicht als HTML



- ▶ Wählen Sie im Menü “Datei” des Modellbrowsers eine der folgenden Optionen:
  - Text exportieren exportiert die Modellübersicht in eine Textdatei.
  - HTML exportieren exportiert die Modellübersicht in eine HTML-Datei.
  - PMML exportieren (nur für IBM DB2 IM-Modelle unterstützt) exportiert das Modell als Predictive Model Markup Language (PMML), das mit anderer PMML-kompatibler Software verwendet werden kann. [Für weitere Informationen siehe Thema Importieren und Exportieren von Modellen als PMML in Kapitel 10 in IBM SPSS Modeler 14.2- Benutzerhandbuch.](#)

*Anmerkung:* Sie können ein generiertes Modell auch mit Knoten speichern im Menü “Datei” speichern. [Für weitere Informationen siehe Thema Durchsuchen von Modell-Nuggets in Kapitel 3 in IBM SPSS Modeler 14.2-Modellierungsknoten.](#)

## Modellkonsistenz

IBM® SPSS® Modeler speichert für jedes generierte Datenbankmodell unter demselben Namen, der in der Datenbank gespeichert ist, eine Beschreibung der Modellstruktur zusammen mit einem Verweis auf das Modell. Auf der Registerkarte “Server” eines generierten Modells wird ein eindeutiger für das Modell erzeugter Schlüssel angezeigt, der mit dem tatsächlichen Modell in der Datenbank übereinstimmt.

Abbildung 2-6  
Generierter Modellschlüssel und Überprüfungsoptionen



Analyseserver-Host:	gb1w2k3dbhost1
Analyseserver-Datenbank:	dan
SQL Server-Verbindung:	SQL Server Stndrd
Modell-GUID:	{2D5DB0DF-5888-43EC-BBC2-1218F057F7AD}

Überprüfen Ansicht

SPSS Modeler überprüft anhand dieses zufällig erstellten Schlüssels, ob das Modell noch konsistent ist. Dieser Schlüssel wird bei der Modellerstellung in der Beschreibung des Modells gespeichert. Es empfiehlt sich, vor der Ausführung eines Bereitstellungs-Streams zu überprüfen, ob die Schlüssel übereinstimmen.

- Klicken Sie auf die Schaltfläche Überprüfen, um durch einen Vergleich der Modellbeschreibung des in der Datenbank gespeicherten Modells mit dem in SPSS Modeler gespeicherten Schlüssel die Konsistenz des Modells zu überprüfen. Wird das Datenbankmodell nicht gefunden oder stimmen die beiden Schlüssel nicht überein, wird ein Fehler ausgegeben.

## Anzeigen und Exportieren von generiertem SQL-Code

Der generierte SQL-Code kann vor der Ausführung angezeigt werden, was für die Fehlersuche hilfreich sein kann. [Für weitere Informationen siehe Thema Vorschau für erzeugte SQL in Kapitel 6 in IBM SPSS Modeler Server 14.2-Verwaltungs- und Leistungshandbuch.](#)

# ***Datenbankmodellierung mit Microsoft Analysis Services***

## ***IBM SPSS Modeler und Microsoft Analysis Services***

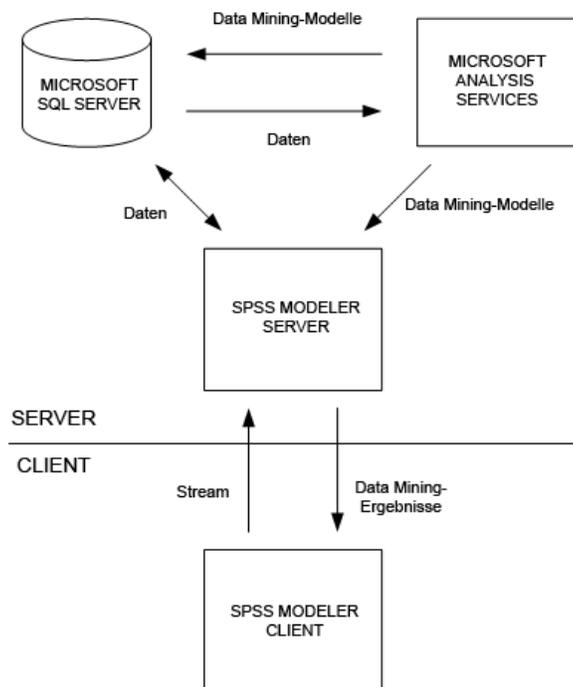
IBM® SPSS® Modeler unterstützt die Integration mit Microsoft SQL Server Analysis Services. Diese Funktionalität ist in SPSS Modeler als Modellierungsknoten implementiert und über die Datenbank-Modellierungspalette verfügbar. Falls die Palette nicht sichtbar ist, aktivieren Sie im Dialogfeld “Hilfsprogramme” auf der Registerkarte “Microsoft” die MS Analysis Services-Integration. [Für weitere Informationen siehe Thema Aktivieren der Integration mit Analysis Services auf S. 15.](#)

SPSS Modeler unterstützt die Integration der folgenden Analysis Services-Algorithmen:

- Decision Trees (Entscheidungsbäume)
- Clusterbildung
- Assoziationsregeln
- Naive Bayes
- Lineare Regression
- Neuronales Netzwerk
- Logistische Regression
- Zeitreihen
- Sequenz-Clustering

Die folgende Abbildung veranschaulicht den Fluss der Daten vom Client zum Server, wobei das In-Database Mining von IBM® SPSS® Modeler Server verwaltet wird. Die Modellerstellung erfolgt mit Analysis Services. Das resultierende Modell wird in Analysis Services gespeichert. Ein Verweis auf dieses Modell bleibt in den SPSS Modeler-Streams erhalten. Das Modell wird dann aus Analysis Services entweder an Microsoft SQL Server oder an SPSS Modeler zum Zwecke des Scoring heruntergeladen.

**Abbildung 3-1**  
 Datenfluss zwischen IBM SPSS Modeler, Microsoft SQL Server und Microsoft Analysis Services bei der Modellerstellung



*Hinweis:* Der SPSS Modeler Server ist nicht erforderlich, kann jedoch verwendet werden. Der IBM® SPSS® Modeler-Client kann In-Database Mining-Berechnungen selbst verarbeiten.

## Anforderungen für die Integration mit Microsoft Analysis Services

Für die Modellerstellung innerhalb der Datenbank unter Verwendung von Analysis Services-Algorithmen mit IBM® SPSS® Modeler gelten die folgenden Voraussetzungen. Wenden Sie sich ggf. an Ihren Datenbankverwalter, um sicherzustellen, dass diese Bedingungen erfüllt sind.

- IBM® SPSS® Modeler wird im Rahmen einer IBM® SPSS® Modeler Server-Installation (verteilter Modus) unter Windows ausgeführt. UNIX-Plattformen werden in dieser Integration mit Analysis Services nicht unterstützt.

*Wichtiger Hinweis:* Die SPSS Modeler-Benutzer müssen eine ODBC-Verbindung mithilfe des SQL Native Client-Treibers herstellen, der unter der unten angegebenen URL unter *Additional SPSS Modeler Server Requirements* (Zusätzliche Anforderungen) bei Microsoft erhältlich ist. Der im Lieferumfang von IBM® SPSS® Data Access Pack enthaltene (und normalerweise für andere Verwendungszwecke von SPSS Modeler empfohlene) Treiber wird hierfür nicht empfohlen. Der Treiber sollte für die Verwendung von SQL Server mit aktivierter Funktion Integrierte Windows-Authentifizierung konfiguriert sein, da SPSS Modeler keine SQL Server-Authentifizierung unterstützt. Wenn Sie Fragen zur Erstellung oder Einstellung von Berechtigungen für ODBC-Datenquellen haben, wenden Sie sich an Ihren Datenbankadministrator.

- SQL Server 2005 oder 2008 muss installiert sein, jedoch nicht unbedingt auf demselben Host wie SPSS Modeler. Die SPSS Modeler müssen über ausreichende Berechtigungen zum Lesen und Schreiben von Daten sowie zum Erstellen und Verwerfen von Tabellen und Ansichten verfügen.  
*Anmerkung:* SQL Server Enterprise Edition wird empfohlen. Die Enterprise Edition bietet zusätzliche Flexibilität durch Bereitstellung erweiterter Parameter zur Feinabstimmung der Algorithmusergebnisse. Die Version Standard Edition enthält dieselben Parameter, der Benutzer kann jedoch einige der erweiterten Parameter nicht bearbeiten.
- Microsoft SQL Server Analysis Services muss auf demselben Host wie SQL Server installiert sein.

### **Weitere IBM SPSS Modeler Server-Anforderungen**

Um Analysis Services-Algorithmen mit SPSS Modeler Server verwenden zu können, müssen folgende Komponenten auf dem SPSS Modeler Server-Hostrechner installiert sein.

*Anmerkung:* Wenn SQL Server auf demselben Host wie SPSS Modeler Server installiert ist, sind diese Komponenten bereits verfügbar.

- Microsoft .NET Framework Version 2.0 Redistributable Package (x86)
- Microsoft Core XML Services (MSXML) 6.0
- Microsoft SQL Server 2008 Analysis Services 10.0 OLE DB Provider (Wählen Sie unbedingt die korrekte Version für Ihr Betriebssystem.)
- Microsoft SQL Server 2008 Native Client (Wählen Sie unbedingt die korrekte Version für Ihr Betriebssystem.)

Um diese Komponenten herunterzuladen, navigieren Sie zu [www.microsoft.com/downloads](http://www.microsoft.com/downloads), suchen Sie .NET Framework bzw. (für alle anderen Komponenten) SQL Server Feature Pack und wählen Sie das neueste Paket für Ihre SQL Server-Version aus.

Möglicherweise müssen zunächst andere Pakete installiert werden. Diese sollten ebenfalls auf der Microsoft Downloads-Website erhältlich sein.

### **Weitere IBM SPSS Modeler-Anforderungen**

Um Analysis Services-Algorithmen mit SPSS Modeler verwenden zu können, müssen dieselben Komponenten installiert sein, wie oben angegeben. Darüber hinaus sind am Client folgende Komponenten erforderlich:

- Microsoft SQL Server 2008 Datamining Viewer Controls (Wählen Sie unbedingt die korrekte Version für Ihr Betriebssystem.) - Dazu ist auch folgende Komponente erforderlich:
- Microsoft ADOMD.NET

Um diese Komponenten herunterzuladen, navigieren Sie zu [www.microsoft.com/downloads](http://www.microsoft.com/downloads), suchen Sie SQL Server Feature Pack und wählen Sie das neueste Paket für Ihre SQL Server-Version aus.

*Hinweis:* Für Datenbankmodellierung und SQL-Optimierung muss auf dem SPSS Modeler-Computer die SPSS Modeler Server-Konnektivität aktiviert sein. Wenn diese Einstellung aktiviert ist, können Sie auf Datenbankalgorithmen zugreifen, SQL direkt aus SPSS Modeler per Pushback übertragen und auf SPSS Modeler Server zugreifen. Um den aktuellen Lizenzstatus zu überprüfen, klicken Sie im Menü “Hilfe” in SPSS Modeler auf Info. [Für weitere Informationen siehe Thema Verbindung mit IBM SPSS Modeler Server in Kapitel 3 in IBM SPSS Modeler 14.2- Benutzerhandbuch.](#)

## **Aktivieren der Integration mit Analysis Services**

Um die Integration von IBM® SPSS® Modeler mit Analysis Services zu ermöglichen, müssen Sie SQL Server und Analysis Services konfigurieren, eine ODBC-Datenquelle erstellen, im SPSS Modeler-Dialogfeld “Hilfsprogramme” die Integration aktivieren und die SQL-Erzeugung und -Optimierung aktivieren.

*Hinweis:* Microsoft SQL Server und Microsoft Analysis Services müssen verfügbar sein. [Für weitere Informationen siehe Thema Anforderungen für die Integration mit Microsoft Analysis Services auf S. 13.](#)

### **Konfigurieren von SQL Server**

Konfigurieren Sie SQL Server so, dass die Möglichkeit des Scoring innerhalb der Datenbank zugelassen wird.

- ▶ Erstellen Sie auf dem SQL Server-Hostcomputer den folgenden Registrierungsschlüssel:

```
HKEY_LOCAL_MACHINE\SOFTWARE\Microsoft\MSSQLServer\Providers\MSOLAP
```

- ▶ Fügen Sie den folgenden DWORD-Wert zu diesem Schlüssel hinzu:

```
AllowInProcess 1
```

- ▶ Starten Sie SQL Server nach dieser Änderung neu.

### **Konfigurieren von Analysis Services**

Bevor SPSS Modeler mit Analysis Services kommunizieren kann, müssen zunächst zwei Einstellungen im Dialogfeld mit den Eigenschaften von Analysis Services manuell konfiguriert werden:

- ▶ Melden Sie sich über MS SQL Server Management Studio beim Analysis Server an.
- ▶ Öffnen Sie das Dialogfeld mit den Eigenschaften. Klicken Sie hierzu mit der rechten Maustaste auf den Servernamen, und wählen Sie die Option Eigenschaften.
- ▶ Aktivieren Sie das Kontrollkästchen Erweiterte (Alle) Eigenschaften anzeigen.

- ▶ Ändern Sie die folgenden Eigenschaften:
  - Ändern Sie den Wert für `DataMining\AllowAdHocOpenRowsetQueries` in `True` (der Standardwert lautet `False`).
  - Ändern Sie den Wert für `DataMining\AllowProvidersInOpenRowset` in `[all]` (hier gibt es keinen Standardwert).

### **Erstellen eines ODBC-DSN für SQL Server**

Um in einer Datenbank zu lesen oder in ihr zu schreiben, muss eine ODBC-Datenquelle für die entsprechende Datenbank mit den erforderlichen Lese- und Schreibberechtigungen installiert und konfiguriert sein. Der Microsoft SQL Native Client ODBC-Treiber ist erforderlich und wird automatisch gemeinsam mit SQL Server installiert. *Der im Lieferumfang von IBM® SPSS® Data Access Pack enthaltene (und normalerweise für andere Verwendungszwecke von SPSS Modeler empfohlene) Treiber wird hierfür nicht empfohlen.* Wenn sich SPSS Modeler und SQL Server auf unterschiedlichen Hosts befinden, können Sie den Microsoft SQL Native Client ODBC-Treiber herunterladen. [Für weitere Informationen siehe Thema Anforderungen für die Integration mit Microsoft Analysis Services auf S. 13.](#)

Wenn Sie Fragen zur Erstellung oder Einstellung von Berechtigungen für ODBC-Datenquellen haben, wenden Sie sich an Ihren Datenbankadministrator.

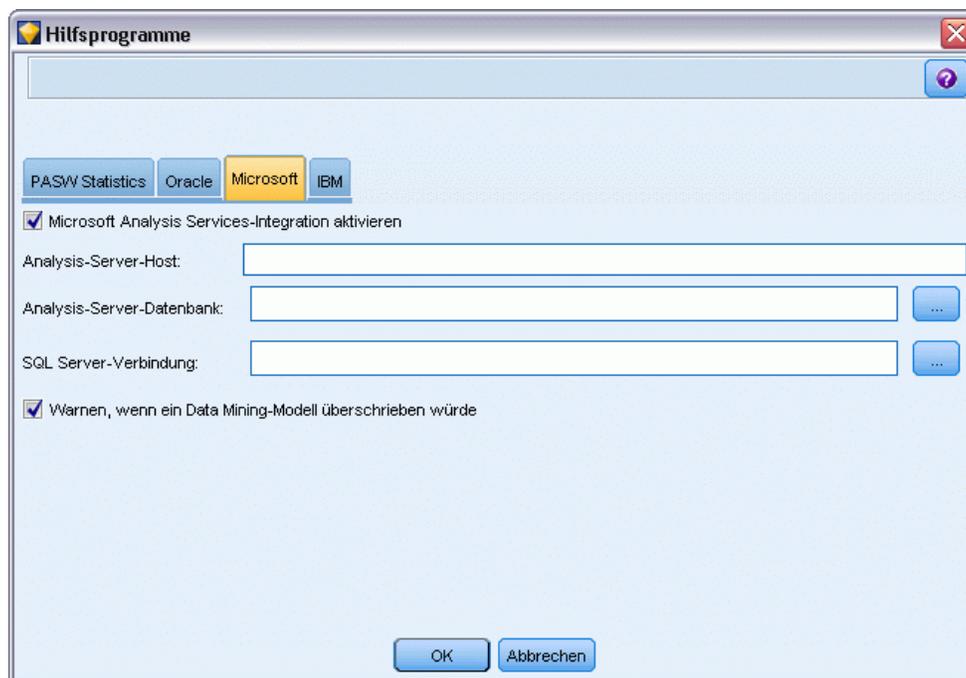
- ▶ Erstellen Sie mit dem Microsoft SQL Native Client ODBC-Treiber einen ODBC DSN, der auf die im Data-Mining-Vorgang verwendete SQL Server-Datenbank verweist. Die restlichen Standardeinstellungen des Treibers sollten unverändert beibehalten werden.
- ▶ Stellen Sie für diesen DSN sicher, dass die Option Integrierte Windows-Authentifizierung aktiviert ist.
  - Wenn IBM® SPSS® Modeler und IBM® SPSS® Modeler Server auf unterschiedlichen Hosts ausgeführt werden, müssen Sie auf beiden Hosts den gleichen ODBC-DSN erstellen. Stellen Sie sicher, dass auf den Hosts jeweils derselbe DSN-Name verwendet wird.

### **Aktivieren der Analysis Services-Integration in IBM SPSS Modeler**

Damit Analysis Services in SPSS Modeler genutzt werden kann, müssen Sie zunächst im Dialogfeld "Hilfsprogramme" einige Angaben zum Server machen.

- ▶ Wählen Sie in den SPSS Modeler-Menüs folgende Befehlsfolge:  
Werkzeuge > Optionen > Hilfsprogramme

Abbildung 3-2  
Dialogfeld "Hilfsprogramme"



- ▶ Klicken Sie auf die Registerkarte Microsoft.
  - **Microsoft Analysis Services-Integration aktivieren.** Wählen Sie diese Option aus, um die Registerkarte "Daten-Modellierung" zur Knotenpalette am unteren Rand des SPSS Modeler-Fensters hinzuzufügen. Mit dieser Registerkarte können Sie auf Analysis Services-Knoten zugreifen (z. B. auf den MS-Entscheidungsbaum-Modellierungsknoten).

Abbildung 3-3  
Registerkarte "Datenbank-Modellierung"



- **Analyseserver-Host.** Geben Sie den Namen des Computers an, auf dem Analysis Services ausgeführt wird.
- **Analyseserver-Datenbank.** Wählen Sie die gewünschte Datenbank aus, indem Sie auf die Schaltfläche mit den Auslassungszeichen (...) klicken. Es wird ein weiteres Dialogfeld geöffnet, in dem Sie eine der verfügbaren Datenbanken auswählen können. In der Liste werden die Datenbanken aufgeführt, die für den angegebenen Analyseserver verfügbar sind. Da Microsoft Analysis Services Data-Mining-Modelle in benannten Datenbanken speichert, sollten Sie die entsprechende Datenbank wählen, in der die mit SPSS Modeler erstellten Microsoft-Modelle gespeichert sind.
- **SQL Server-Verbindung** Geben Sie die DSN-Informationen an, die von der SQL Server-Datenbank zum Speichern der Daten verwendet werden, die an den Analyseserver weitergeleitet werden sollen. Wählen Sie die ODBC-Datenquelle, aus der die Daten für

die Erstellung von Analysis Services Data-Mining-Modellen bereitgestellt werden. Wenn Sie Analysis Services-Modelle aus Daten von Einfachdateien oder ODBC-Datenquellen erstellen, werden die Daten automatisch in eine temporäre Tabelle hochgeladen, die in der SQL Server-Datenbank erstellt wird, auf die diese ODBC-Datenquelle verweist.

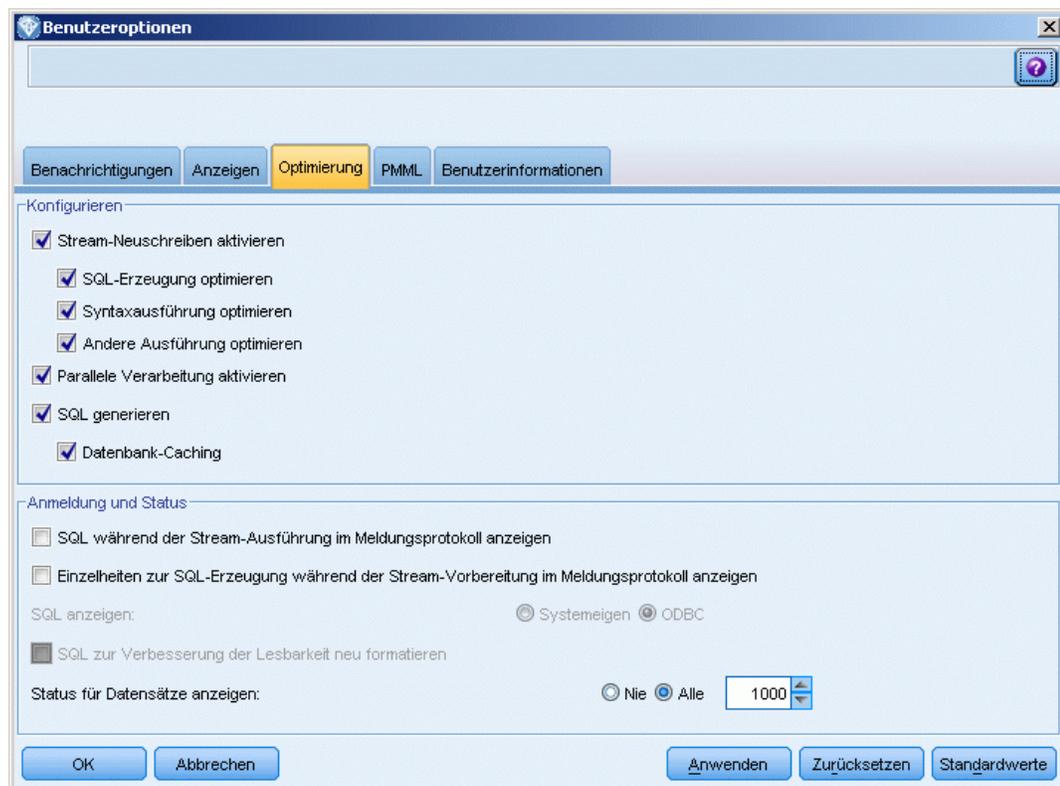
- **Warnen, wenn ein Data-Mining-Modell überschrieben würde.** Wählen Sie diese Option, um sicherzustellen, dass in der Datenbank gespeicherte Modelle nicht von SPSS Modeler überschrieben werden, ohne dass eine Warnung ausgegeben wird.

*Hinweis:* Im Dialogfeld “Hilfsprogramme” vorgenommene Einstellungen können innerhalb der verschiedenen Analysis Services-Knoten überschrieben werden.

### **Aktivieren der SQL-Erzeugung und -Optimierung**

- ▶ Wählen Sie in den SPSS Modeler-Menüs folgende Befehlsfolge:  
Werkzeuge > Optionen > Benutzeroptionen

Abbildung 3-4  
Optimierungseinstellungen



- ▶ Klicken Sie dann auf die Registerkarte Optimierung.
- ▶ Überzeugen Sie sich, dass die Option SQL generieren aktiviert ist. Diese Einstellung ist für die Datenbank-Modellierung erforderlich.
- ▶ Wählen Sie SQL-Erzeugung optimieren und Andere Ausführung optimieren aus. (Diese Einstellungen sind nicht unbedingt erforderlich, werden aber zur Leistungsoptimierung empfohlen.)

Für weitere Informationen siehe Thema Festlegen von Optimierungsoptionen in Kapitel 12 in *IBM SPSS Modeler 14.2- Benutzerhandbuch*.

## Erstellen von Modellen mit Analysis Services

Für die Modellbildung von Analysis Services muss sich das Trainings-Daten-Set in einer Tabelle oder Ansicht innerhalb der SQL Server-Datenbank befinden. Wenn sich die Daten nicht in SQL Server befinden oder wenn die Daten wegen einer Datenvorbereitung, die nicht in SQL Server erfolgen kann, in IBM® SPSS® Modeler verarbeitet werden müssen, werden diese vor der Modellbildung automatisch in eine temporäre SQL Server-Tabelle geladen.

## Verwalten von Analysis Services-Modellen

Bei der Bildung eines Analysis Services-Modells mit IBM® SPSS® Modeler wird in SPSS Modeler ein Modell erzeugt und außerdem in der SQL Server-Datenbank ein Modell erzeugt oder ersetzt. Das SPSS Modeler-Modell stellt einen Bezug zum Inhalt eines in einem Datenbankserver gespeicherten Datenbankmodells her. SPSS Modeler kann eine Konsistenzprüfung durchführen, indem sowohl im SPSS Modeler-Modell als auch im SQL Server-Modell eine identische, generierte Modellschlüsselzeichenkette gespeichert wird.



Der MS-Modellierungsknoten **Entscheidungsbaum** wird für Vorhersagemodelle mit kategorialen und kontinuierlichen Attributen verwendet. Bei kategorialen Attributen erstellt der Knoten Vorhersagen auf der Grundlage der Beziehungen zwischen den Eingabespalten in einem Daten-Set. Beispiel: Wenn prognostiziert werden soll, welche Kunden mit hoher Wahrscheinlichkeit ein Fahrrad kaufen und neun von zehn jüngeren Kunden ein Fahrrad kaufen, jedoch nur zwei von zehn älteren Kunden, folgert der Knoten, dass das Alter ein guter Prädiktor für den Fahrradkauf ist. Der Entscheidungsbaum erstellt seine Vorhersagen dann auf der Grundlage dieser Tendenz hin zu einem bestimmten Ergebnis. Bei kontinuierlichen Attributen verwendet der Algorithmus lineare Regression, um zu ermitteln, an welcher Stelle sich ein Entscheidungsbaum aufspaltet. Wenn mehrere Spalten auf "vorhersagbar" gesetzt sind oder wenn die Eingangsdaten eine verschachtelte Tabelle enthalten, die auf "vorhersagbar" gesetzt ist, erstellt der Knoten einen gesonderten Entscheidungsbaum für jede vorhersagbare Spalte.



Der **MS Clustering**-Modellierungsknoten verwendet iterative Verfahren zur Gruppierung von Fällen in einem Datenset in Clustern, die ähnliche Merkmale enthalten. Diese Gruppierungen sind sinnvoll für die Untersuchung von Daten, die Identifizierung von Anomalien in den Daten und die Erstellung von Vorhersagen. Clustermodelle identifizieren Beziehungen in einem Datenset, die sich möglicherweise nicht logisch durch Fallbeobachtungen ableiten lassen. Sie können beispielsweise durch Logik feststellen, dass Personen, die mit dem Fahrrad zum Arbeitsplatz pendeln, normalerweise nicht sonderlich weit von ihrem Arbeitsplatz entfernt wohnen. Der Algorithmus kann jedoch noch andere Merkmale von Fahrradpendlern ermitteln, die nicht so offensichtlich sind. Der Clusterknoten unterscheidet sich darin von anderen Data-Mining-Knoten, dass kein Zielfeld angegeben ist. Der Clusterknoten trainiert das Modell ausschließlich ausgehend von den Beziehungen, die in den Daten vorliegen, und von den Clustern, die der Knoten identifiziert.



Der MS Modellierungsknoten **Assoziationsregeln** ist nützlich für Empfehlungs-Engines. Eine Empfehlungs-Engine empfiehlt Kunden Produkte auf der Grundlage der Artikel, die sie bereits erworben oder an denen sie Interesse bekundet haben. Assoziationsmodelle werden für Daten-Sets erstellt, die IDs sowohl für die einzelnen Fälle aufweisen als auch für die Elemente, die diese Fälle enthalten. Eine Gruppe von Elementen in einem Fall wird als **Elementsatz** bezeichnet. Assoziationsmodelle bestehen aus einer Reihe von Elementsätzen und den Regeln, die beschreiben, wie diese Elemente innerhalb der Fälle in Gruppen zusammengefasst werden. Die Regeln, die der Algorithmus ermittelt, können verwendet werden, um anhand der Artikel, die sich bereits im Einkaufswagen des Kunden befinden, vorherzusagen, welche Artikel er voraussichtlich in der Zukunft erwerben wird.



Der MS-Modellierungsknoten **Naive Bayes** von Analysis Services berechnet die bedingte Wahrscheinlichkeit zwischen Ziel- und Prädiktorfeldern und geht dabei davon aus, dass die Spalten unabhängig sind. Das Modell wird als “naiv” bezeichnet, weil es alle vorgeschlagenen Vorhersagevariablen als voneinander unabhängig behandelt. Diese Methode erfordert weniger Berechnungsaufwand als die anderen Analysis Services-Algorithmen und ist daher nützlich für die schnelle Ermittlung von Beziehungen in den vorbereitenden Phasen der Modellierung. Mit diesem Knoten können Sie erste Untersuchungen der Daten vornehmen und anschließend die Ergebnisse anwenden, um zusätzliche Modelle mit anderen Knoten zu erstellen, deren Berechnung länger dauert, die jedoch zu genaueren Ergebnissen führen.



Der MS-Modellierungsknoten **Lineare Regression** ist eine Abwandlung des Knotens “Entscheidungsbäume”, bei dem der Parameter `MINIMUM_LEAF_CASES` auf größer oder gleich der Gesamtzahl der Fälle im Daten-Set gesetzt ist, die der Knoten für das Trainieren des Mining-Modells verwendet. Wenn der Parameter so gesetzt ist, erstellt der Knoten nie eine Aufteilung und führt also eine lineare Regression durch.



Der MS-Modellierungsknoten **Neuronales Netzwerk** ähnelt dem MS-Knoten “Entscheidungsbäume” dahingehend, dass der MS-Knoten “Neuronales Netzwerk” Wahrscheinlichkeiten für jeden möglichen Status des Eingabeattributs berechnet, wenn jeder Status des vorhersagbaren Attributs vorliegt. Später können Sie mithilfe dieser Wahrscheinlichkeiten auf der Grundlage der Eingabeattribute ein Ergebnis des vorhergesagten Attributs prognostizieren.



Der MS-Modellierungsknoten **Logistische Regression** ist eine Abwandlung des MS-Knotens “Neuronales Netzwerk”, bei dem der Parameter `HIDDEN_NODE_RATIO` auf 0 gesetzt ist. Diese Einstellung erstellt ein neuronales Netzwerkmodell, das keine verdeckte Schicht enthält und daher der logistischen Regression entspricht.



Der MS Modellierungsknoten **Zeitreihen** bietet Regressionsalgorithmen, die zur Prognose von stetigen Werten wie z. B. Produktverkäufen im Laufe der Zeit optimiert sind. Während andere Microsoft-Algorithmen, z. B. Entscheidungsbäume, zusätzliche Spalten mit neuen Informationen erfordern, um einen Trend vorherzusagen, verzichtet ein Zeitreihenmodell darauf. Ein Zeitreihenmodell kann Trends allein auf der Basis des ursprünglichen Datensets vorhersagen, mit dem das Modell erstellt wurde. Sie können dem Modell auch neue Daten hinzufügen, wenn Sie eine Vorhersage treffen, und automatisch die neuen Daten in der Trendanalyse berücksichtigen. [Für weitere Informationen siehe Thema MS Time Series-Knoten auf S. 30.](#)



Der MS-Modellierungsknoten **Sequenz-Clustering** identifiziert geordnete Sequenzen in Daten und kombiniert die Ergebnisse dieser Analyse mit Clustering-Techniken, um Cluster auf der Grundlage der Sequenzen und anderer Attribute zu generieren. [Für weitere Informationen siehe Thema MS-Sequenz-Clustering-Knoten auf S. 34.](#)

Sie können von der Datenbank-Modellierungspalette am unteren Rand des SPSS Modeler-Fensters aus auf alle Knoten zugreifen.

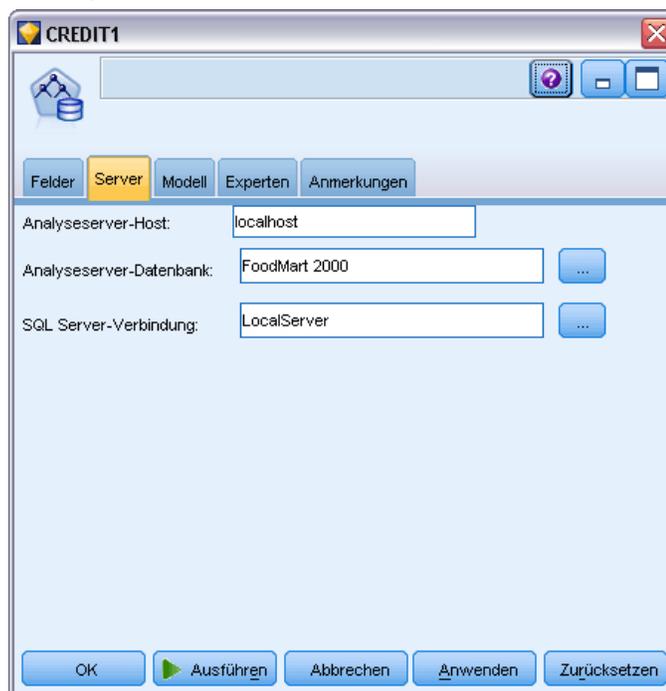
## **Gemeinsame Einstellungen für alle Algorithmenknoten**

Folgende Einstellungen haben alle Analysis Services-Algorithmen gemeinsam.

### **Serveroptionen**

Auf der Registerkarte “Server” können Sie den Analyseserver-Host und die Analyseserver-Datenbank sowie die SQL Server-Datenquelle konfigurieren. Die hier festgelegten Optionen überschreiben die Optionen, die auf der Registerkarte “Microsoft” im Dialogfeld “Hilfsprogramme” festgelegt wurden. [Für weitere Informationen siehe Thema Aktivieren der Integration mit Analysis Services auf S. 15.](#)

Abbildung 3-5  
Serveroptionen

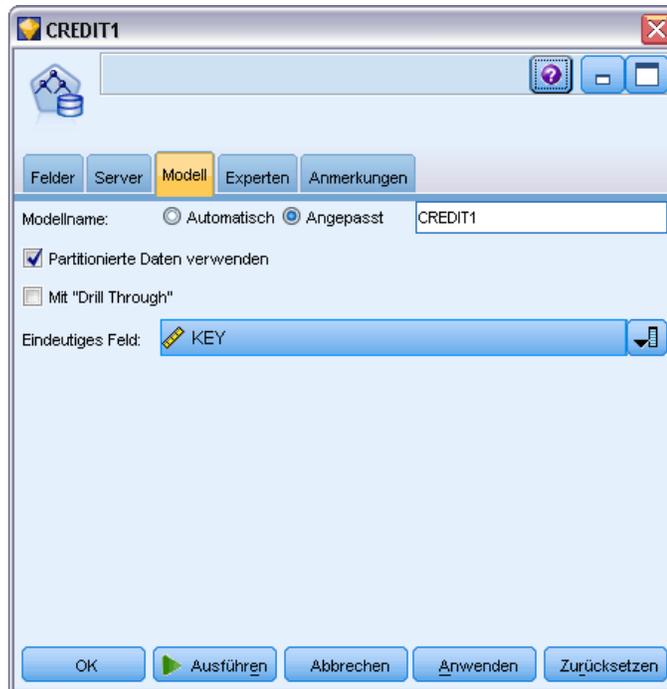


*Hinweis:* Beim Scoring von Analysis Services-Modellen steht eine Variante dieser Registerkarte zur Verfügung. [Für weitere Informationen siehe Thema Analysis Services-Modell-Nugget – Registerkarte “Server” auf S. 37.](#)

### **Modelloptionen**

Um das grundlegendste Modell erstellen zu können, müssen Sie auf der Registerkarte “Modell” Optionen festlegen, bevor Sie weitere Schritte durchführen. Die Scoring-Methode und andere erweiterte Optionen werden auf der Registerkarte “Experten” festgelegt.

Abbildung 3-6  
Modelloptionen



Die folgenden grundlegenden Modellierungsoptionen sind verfügbar:

**Modellname.** Gibt den Namen des Modells an, das beim Ausführen des Knotens erstellt wird.

- **Auto.** Generiert den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen oder dem Namen des Modelltyps in Fällen, in denen kein Ziel angegeben ist (z. B. Clustermodelle).
- **Benutzerdefiniert.** Hier können Sie einen benutzerdefinierten Namen für das erstellte Modell angeben.

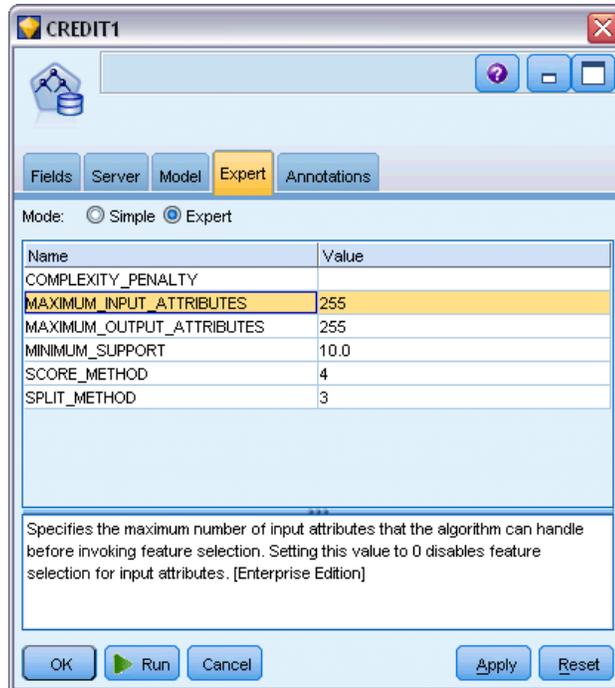
**Partitionierte Daten verwenden.** Teilt die Daten in separate Untergruppen oder Stichproben für das Training, Testen und die Validierung, basierend auf dem aktuellen Partitionsfeld auf. Indem Sie mit einer Stichprobe das Modell erstellen und es mit einer separaten Stichprobe testen, erhalten Sie einen guten Hinweis dafür, wie gut das Modell sich für größere Datenmengen generalisieren lässt, die den aktuellen Daten ähneln. Wenn im Stream kein Partitionsfeld angegeben ist, wird diese Option ignoriert. [Für weitere Informationen siehe Thema Partitionsknoten in Kapitel 4 in IBM SPSS Modeler 14.2- Quellen-, Prozess- und Ausgabeknoten.](#)

**Mit Drillthrough.** Wenn angezeigt, können Sie mithilfe dieser Option das Modell abfragen, um Einzelheiten über die darin enthaltenen Fälle zu erfahren.

**Eindeutiges Feld.** Wählen Sie aus der Dropdownliste ein Feld aus, das jeden Fall eindeutig identifiziert. Im Normalfall ist dies ein ID-Feld, wie z. B. CustomerID.

## MS Expertenoptionen für Entscheidungsbäume

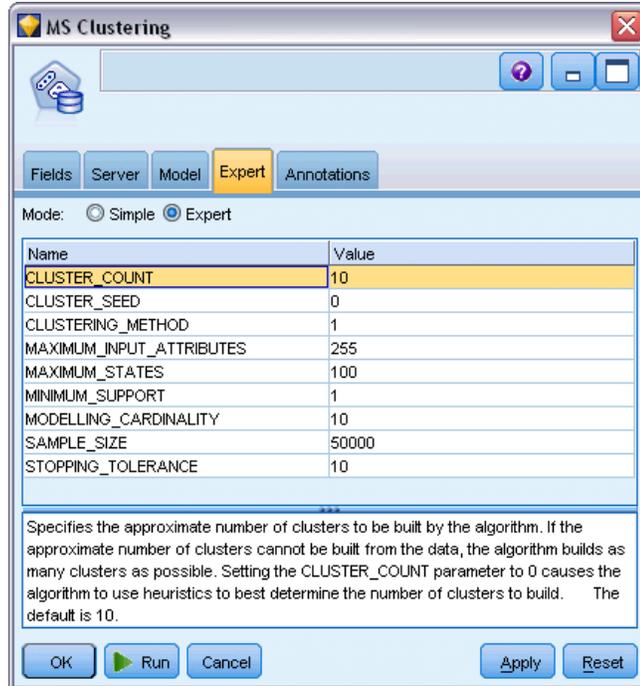
Abbildung 3-7  
MS Expertenoptionen für Entscheidungsbäume



Die auf der Registerkarte "Experten" verfügbaren Optionen können je nach der Struktur des ausgewählten Streams variieren. Vollständige Details zu den für den ausgewählten Modellknoten von Analysis Server verfügbaren Expertenoptionen finden sie in der Hilfe zu den einzelnen Feldern in der Benutzeroberfläche.

## MS Expertenoptionen für Clusterbildung

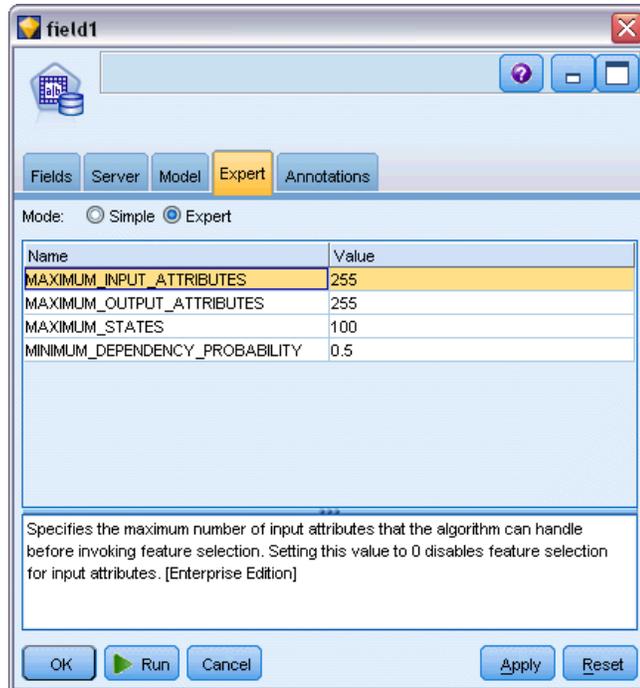
Abbildung 3-8  
MS Expertenoptionen für Clusterbildung



Die auf der Registerkarte "Experten" verfügbaren Optionen können je nach der Struktur des ausgewählten Streams variieren. Vollständige Details zu den für den ausgewählten Modellknoten von Analysis Server verfügbaren Expertenoptionen finden sie in der Hilfe zu den einzelnen Feldern in der Benutzeroberfläche.

## MS Expertenoptionen für Naive Bayes

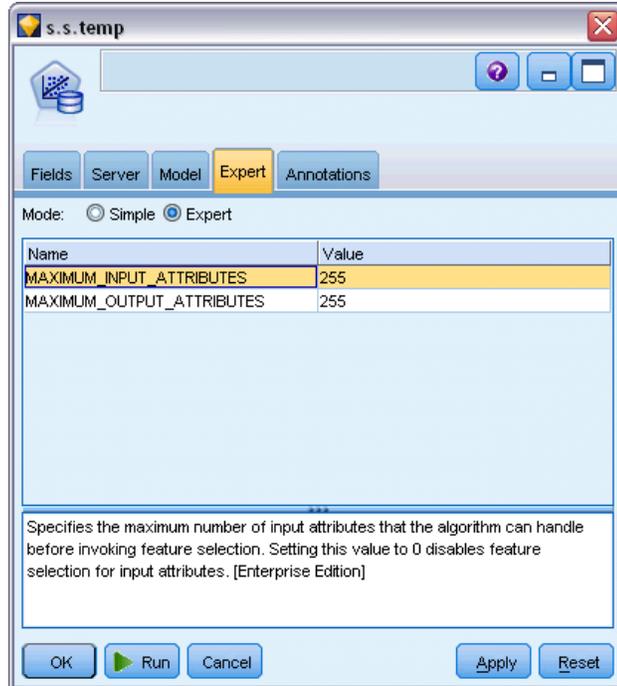
Abbildung 3-9  
MS Expertenoptionen für Naive Bayes



Die auf der Registerkarte "Experten" verfügbaren Optionen können je nach der Struktur des ausgewählten Streams variieren. Vollständige Details zu den für den ausgewählten Modellknoten von Analysis Server verfügbaren Expertenoptionen finden sie in der Hilfe zu den einzelnen Feldern in der Benutzeroberfläche.

## MS Lineare Regression – Expertenoptionen

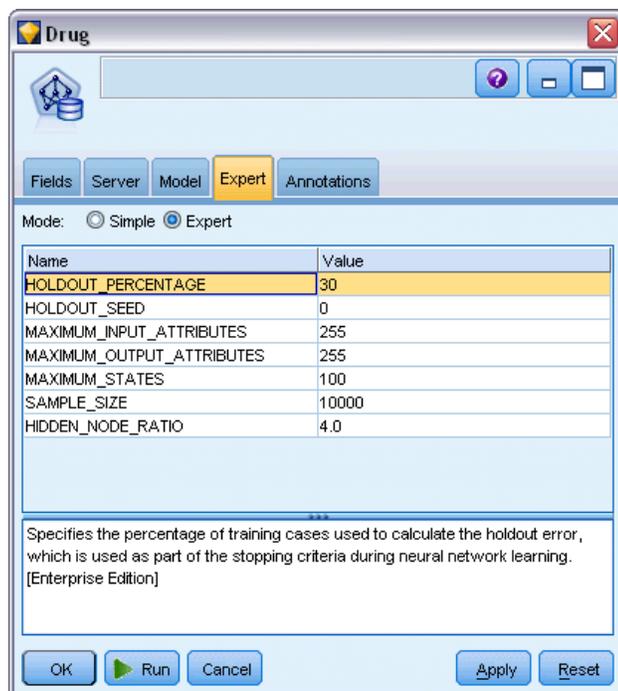
Abbildung 3-10  
MS Lineare Regression – Expertenoptionen



Die auf der Registerkarte "Experten" verfügbaren Optionen können je nach der Struktur des ausgewählten Streams variieren. Vollständige Details zu den für den ausgewählten Modellknoten von Analysis Server verfügbaren Expertenoptionen finden sie in der Hilfe zu den einzelnen Feldern in der Benutzeroberfläche.

## MS Neuronales Netzwerk – Expertenoptionen

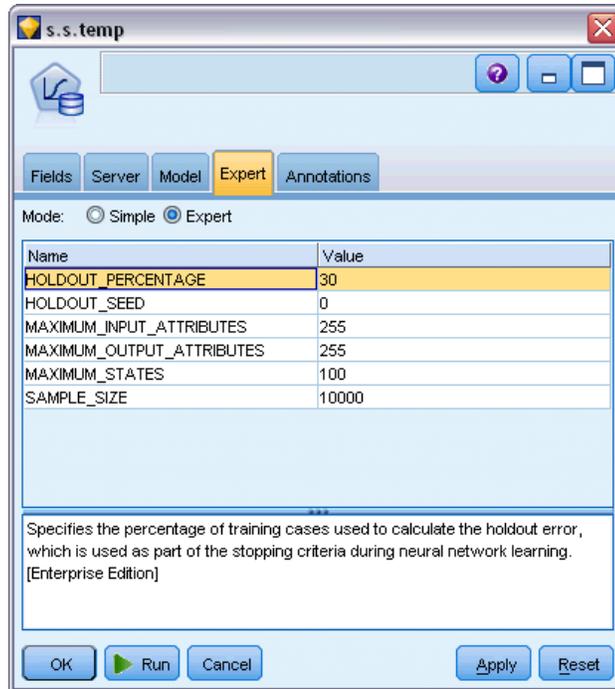
Abbildung 3-11  
MS Neuronales Netzwerk – Expertenoptionen



Die auf der Registerkarte “Experten” verfügbaren Optionen können je nach der Struktur des ausgewählten Streams variieren. Vollständige Details zu den für den ausgewählten Modellknoten von Analysis Server verfügbaren Expertenoptionen finden sie in der Hilfe zu den einzelnen Feldern in der Benutzeroberfläche.

## MS Logistische Regression – Expertenoptionen

Abbildung 3-12  
MS Logistische Regression – Expertenoptionen



Die auf der Registerkarte “Experten” verfügbaren Optionen können je nach der Struktur des ausgewählten Streams variieren. Vollständige Details zu den für den ausgewählten Modellknoten von Analysis Server verfügbaren Expertenoptionen finden sie in der Hilfe zu den einzelnen Feldern in der Benutzeroberfläche.

## MS-Assoziationsregel-Knoten

Der MS Modellierungsknoten “Assoziationsregeln” ist nützlich für Empfehlungs-Engines. Eine Empfehlungs-Engine empfiehlt Kunden Produkte auf der Grundlage der Artikel, die sie bereits erworben oder an denen sie Interesse bekundet haben. Assoziationsmodelle werden für Daten-Sets erstellt, die IDs sowohl für die einzelnen Fälle aufweisen als auch für die Elemente, die diese Fälle enthalten. Eine Gruppe von Elementen in einem Fall wird als **Elementsatz** bezeichnet.

Assoziationsmodelle bestehen aus einer Reihe von Elementsätzen und den Regeln, die beschreiben, wie diese Elemente innerhalb der Fälle in Gruppen zusammengefasst werden. Die Regeln, die der Algorithmus ermittelt, können verwendet werden, um anhand der Artikel, die sich bereits im Einkaufswagen des Kunden befinden, vorherzusagen, welche Artikel er voraussichtlich in der Zukunft erwerben wird.

Für Daten in Tabellenformat erzeugt der Algorithmus Werte zur Darstellung der Wahrscheinlichkeit (*\$MP-Feld*) für jede generierte Empfehlung (*\$M-Feld*). Für Daten in Transaktionsformat werden Werte für Unterstützung (*\$MS-Feld*), Wahrscheinlichkeit (*\$MP-Feld*)

und angepasste Wahrscheinlichkeit (\$MAP-Feld) für jede generierte Empfehlung (\$M-Feld) erstellt. Für weitere Informationen siehe Thema [Tabellendaten im Vergleich zu Transaktionsdaten](#) in Kapitel 12 in *IBM SPSS Modeler 14.2-Modellierungsknoten*.

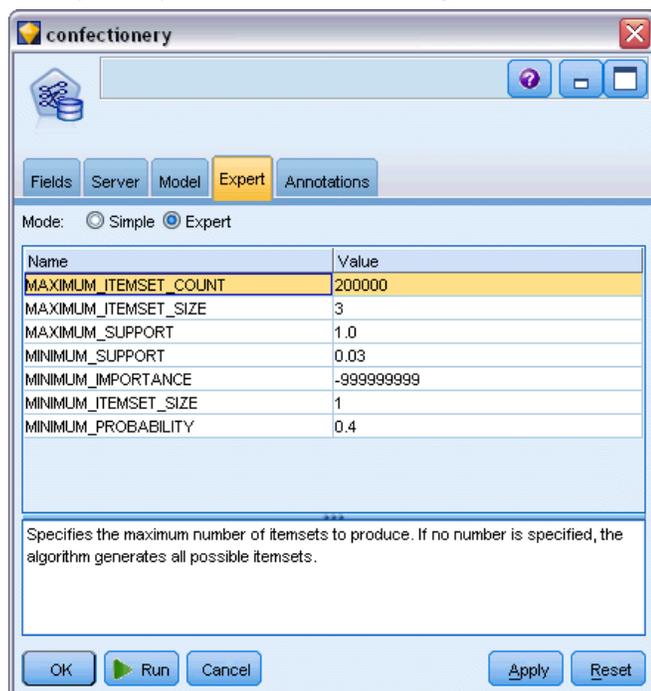
### Voraussetzungen

Die Anforderungen für ein transaktionelles Assoziationsmodell sehen wie folgt aus:

- **Eindeutiges Feld.** Ein Assoziationsregelmodell erfordert einen Schlüssel, der Datensätze eindeutig identifiziert.
- **ID-Feld.** Beim Aufbau eines MS Assoziationsregelmodells mit Daten in Transaktionsformat ist ein ID-Feld erforderlich, das jede Transaktion identifiziert. ID-Felder können auf dieselben Werte gesetzt werden wie das eindeutige Feld.
- **Mindestens ein Eingabefeld.** Der Assoziationsregel-Algorithmus verlangt mindestens ein Eingabefeld.
- **Zielfeld.** Beim Erstellen eines MS Assoziationsmodells mit Transaktionsdaten muss das Zielfeld das Transaktionsfeld sein, z. B. Produkte, die ein Benutzer gekauft hat.

### MS Expertenoptionen für Assoziationsregeln

Abbildung 3-13  
MS Expertenoptionen für Assoziationsregeln



Die auf der Registerkarte "Experten" verfügbaren Optionen können je nach der Struktur des ausgewählten Streams variieren. Vollständige Details zu den für den ausgewählten Modellknoten von Analysis Server verfügbaren Expertenoptionen finden sie in der Hilfe zu den einzelnen Feldern in der Benutzeroberfläche.

## **MS Time Series-Knoten**

Der MS Time Series-Modellierungsknoten unterstützt zwei Arten der Vorhersage:

- Zukunft
- Historisch

**Zukunftsvorhersagen** schätzen Zielfeldwerte für eine angegebene Anzahl an Zeitspannen über das Ende Ihrer historischen Daten hinaus und werden immer ausgeführt. **Historische Vorhersagen** sind geschätzte Zielfeldwerte für eine angegebene Anzahl an Zeitspannen, für die Ihre historischen Daten die tatsächlichen Werte enthalten. Sie können mithilfe historischer Vorhersagen die Qualität des Modells beurteilen, indem Sie die tatsächlichen historischen Werte mit den vorhergesagten Werten vergleichen. Der Wert des Anfangspunkts für die Vorhersagen bestimmt, ob historische Vorhersagen ausgeführt werden.

Im Unterschied zum IBM® SPSS® Modeler-Zeitreihenknoten benötigt der MS Time Series-Knoten keinen vorangehenden Zeitintervallknoten. Ein weiterer Unterschied besteht darin, dass Werte standardmäßig nur für die vorhergesagten Zeilen erzeugt werden, nicht für alle historischen Zeilen in den Zeitreihendaten.

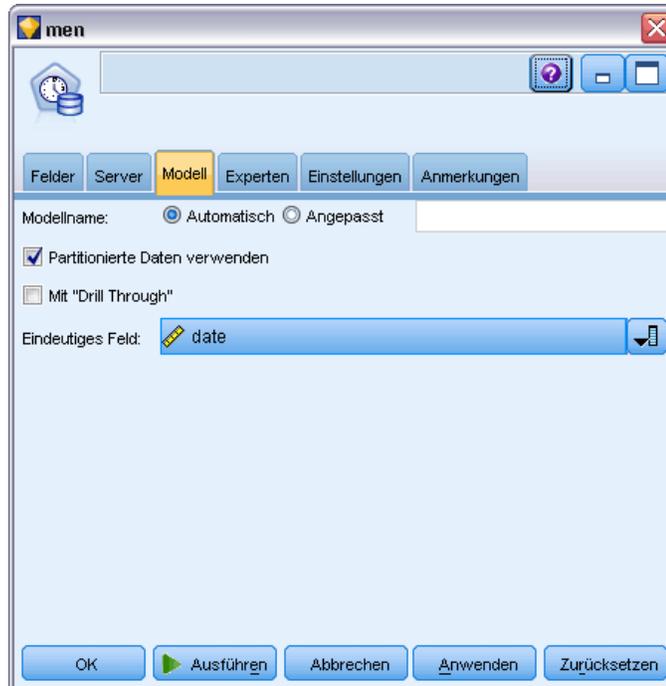
### **Voraussetzungen**

Die Anforderungen für ein MS Time Series-Modell sehen wie folgt aus:

- **Einzelnes Schlüsselzeitfeld.** Jedes Modell muss ein Zahlen- oder Datumsfeld enthalten, das als die Fallreihe verwendet wird und das Zeitintervall definiert, welches das Modell verwendet. Der Datentyp für das Schlüsselzeitfeld kann entweder Datum/Uhrzeit oder Zahl sein. Jedoch muss das Feld bestimmte stetige Werte enthalten, die für jede Reihe eindeutig sein müssen.
- **Einzelnes Zielfeld.** Sie können in jedem Modell nur ein Zielfeld angeben. Der Datentyp des Zielfelds muss stetige Werte haben. Sie können beispielsweise vorhersagen, wie numerische Attribute wie Einnahmen, Umsatz oder Temperatur sich im Laufe der Zeit ändern. Jedoch können Sie kein Feld verwenden, das kategoriale Werte als Zielfeld enthält, z. B. Kaufstatus oder Bildungsniveau.
- **Mindestens ein Eingabefeld.** Der MS Time Series-Algorithmus verlangt mindestens ein Eingabefeld. Der Datentyp des Eingabefelds muss stetige Werte haben. Nicht stetige Eingabefelder werden bei der Erstellung des Modells ignoriert.
- **Datenset muss sortiert sein.** Das Eingabe-Datenset muss (nach dem Schlüsselzeitfeld) sortiert sein, ansonsten wird die Modellerstellung mit einem Fehler abgebrochen.

## MS Zeitreihenmodelle – Optionen

Abbildung 3-14  
MS Zeitreihenmodelle – Optionen



**Modellname.** Gibt den Namen des Modells an, das beim Ausführen des Knotens erstellt wird.

- **Auto.** Generiert den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen oder dem Namen des Modelltyps in Fällen, in denen kein Ziel angegeben ist (z. B. Clustermodelle).
- **Benutzerdefiniert.** Hier können Sie einen benutzerdefinierten Namen für das erstellte Modell angeben.

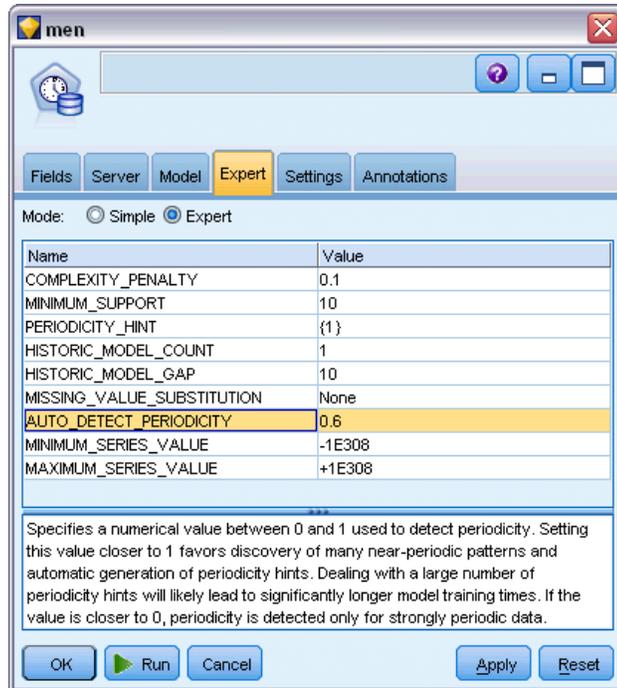
**Partitionierte Daten verwenden.** Wenn ein Partitionsfeld definiert ist, gewährleistet diese Option, dass nur Daten aus der Trainingspartition für die Modellerstellung verwendet werden. [Für weitere Informationen siehe Thema Partitionsknoten in Kapitel 4 in IBM SPSS Modeler 14.2- Quellen-, Prozess- und Ausgabeknoten.](#)

**Mit Drillthrough.** Wenn angezeigt, können Sie mithilfe dieser Option das Modell abfragen, um Einzelheiten über die darin enthaltenen Fälle zu erfahren.

**Eindeutiges Feld.** Wählen Sie aus der Dropdown-Liste das Schlüsselzeitfeld, das zur Erstellung des Zeitreihenmodells verwendet wird.

### MS Zeitreihen – Expertenoptionen

Abbildung 3-15  
MS Zeitreihen – Expertenoptionen

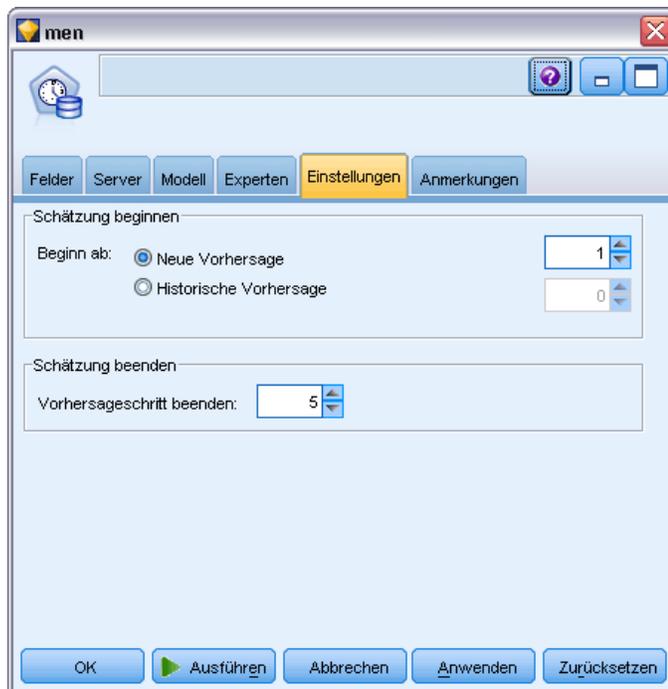


Die auf der Registerkarte “Experten” verfügbaren Optionen können je nach der Struktur des ausgewählten Streams variieren. Vollständige Details zu den für den ausgewählten Modellknoten von Analysis Server verfügbaren Expertenoptionen finden sie in der Hilfe zu den einzelnen Feldern in der Benutzeroberfläche.

Für historische Vorhersagen entscheidet sich die Anzahl der historischen Schritte, die im Scoring-Ergebnis berücksichtigt werden können, durch den Wert von  $(\text{HISTORIC\_MODEL\_COUNT} * \text{HISTORIC\_MODEL\_GAP})$ . Standardmäßig beträgt die Begrenzung 10, d. h., nur zehn historische Vorhersagen werden getroffen. In diesem Fall tritt z. B. ein Fehler auf, wenn Sie einen Wert kleiner als -10 für Historische Vorhersagen in der Registerkarte “Einstellungen” des Modell-Nuggets eingeben (siehe [MS Zeitreihen-Modell-Nugget – Registerkarte “Einstellungen” auf S. 43](#)). Wenn Sie mehr historische Vorhersagen sehen möchten, können Sie den Wert von HISTORIC\_MODEL\_COUNT oder HISTORIC\_MODEL\_GAP erhöhen, wodurch sich allerdings auch die Erstellungsdauer für das Modell verlängert.

## MS Zeitreihen – Einstellungsoptionen

Abbildung 3-16  
MS Zeitreihen – Einstellungsoptionen



**Schätzung beginnen.** Geben Sie die Zeitperiode an, in der Vorhersagen beginnen sollen.

- **Start: Neue Vorhersage.** Die Zeitperiode, in der zukünftige Vorhersagen beginnen sollen, ausgedrückt als Versatz von der letzten Zeitperiode Ihrer historischen Daten. Beispiel: Wenn Ihre historischen Daten bei 12.99 enden und Ihre Vorhersagen 01.00 beginnen sollen, verwenden Sie den Wert 1. Wenn die Vorhersagen jedoch 03.00 beginnen sollen, verwenden Sie den Wert 3.
- **Start: Historische Vorhersage.** Die Zeitperiode, in der historische Vorhersagen beginnen sollen, ausgedrückt als negativer Versatz von der letzten Zeitperiode Ihrer historischen Daten. Beispiel: Wenn Ihre historischen Daten bei 12.99 enden und Sie historische Vorhersagen für die letzten fünf Zeitperioden Ihrer Daten erstellen wollen, verwenden Sie den Wert -5.

**Schätzung beenden.** Geben Sie die Zeitperiode an, in der Vorhersagen enden sollen.

- **Letzter Schritt der Vorhersage.** Die Zeitperiode, in der zukünftige Vorhersagen enden sollen, ausgedrückt als Versatz von der letzten Zeitperiode Ihrer historischen Daten. Beispiel: Wenn Ihre historischen Daten bei 12.99 enden und Ihre Vorhersagen 6.00 enden sollen, verwenden Sie hier den Wert 6. Für zukünftige Vorhersagen muss der Wert stets größer oder gleich dem Start-Wert sein.

## **MS-Sequenz-Clustering-Knoten**

Der MS-Sequenz-Clustering-Knoten verwendet einen Sequenzanalyse-Algorithmus zur Untersuchung von Daten, die Ereignisse enthalten, die durch nachfolgende Pfade bzw. *Sequenzen* verknüpft werden können. Einige Beispiele dafür können die Klickpfade sein, die angelegt werden, wenn Benutzer in einer Website navigieren oder suchen, oder die Reihenfolge, in der ein Kunde Artikel in seinen Einkaufswagen bei einem Online-Händler legt. Der Algorithmus findet die häufigsten Sequenzen durch Gruppierung bzw. *Clustering* von Sequenzen, die identisch sind.

### **Voraussetzungen**

Die Anforderungen für ein Microsoft Sequenz-Clustering-Modell sehen wie folgt aus:

- **ID-Feld.** Der Microsoft Sequenz-Clustering-Algorithmus verlangt, dass die Sequenzinformationen in Transaktionsformat gespeichert sind (siehe [Tabellendaten im Vergleich zu Transaktionsdaten auf S.](#) ). Dafür ist ein ID-Feld erforderlich, das jede Transaktion identifiziert.
- **Mindestens ein Eingabefeld.** Der Algorithmus verlangt mindestens ein Eingabefeld.
- **Sequenzfeld.** Der Algorithmus erfordert auch ein Sequenz-ID-Feld mit einem Messniveau des Typs “Stetig”. Sie können beispielsweise eine Webseiten-ID, eine Ganzzahl oder eine Zeichenkette verwenden, solange das Feld Ereignisse in einer Sequenz identifiziert. Nur eine Sequenz-ID ist pro Sequenz zulässig und nur ein Sequenztyp ist pro Modell erlaubt. Das Sequenzfeld muss sich von den Feldern “ID” und “Eindeutig” unterscheiden.
- **Zielfeld.** Ein Zielfeld ist beim Erstellen eines Sequenz-Clustering-Modells erforderlich.
- **Eindeutiges Feld.** Ein Sequenz-Clustering-Modell erfordert ein Schlüsselfeld, das Datensätze eindeutig identifiziert. Sie können das Feld “Eindeutig” auf denselben Wert setzen wie das Feld “ID”.

### **MS Sequenz-Clustering - Feldoptionen**

Alle Modellierungsknoten besitzen die Registerkarte “Felder”, auf der Sie die Felder festlegen, die beim Erstellen des Modells verwendet werden.

Abbildung 3-17  
Angaben der Felder für MS Sequenz-Clustering



Bevor Sie ein Sequenz-Clustering-Modell erstellen können, müssen Sie festlegen, welche Felder als Ziele und als Eingaben verwendet werden sollen. Beachten Sie, dass Sie für den MS-Sequenz-Clustering-Knoten keine Feldinformationen aus einem Typenknoten weiter oben im Stream verwenden können. Sie müssen die Feldeinstellungen hier angeben.

**ID.** Wählen Sie ein ID-Feld aus der Liste aus. Als ID-Feld können numerische oder symbolische Felder verwendet werden. Jeder eindeutige Wert in diesem Feld sollte eine bestimmte Analyseeinheit darstellen. Bei einer Warenkorbanwendung könnte z. B. jede ID einen einzelnen Kunden darstellen. Für eine Webprotokoll-Analyseanwendung könnte jede ID einen Computer (nach IP-Adresse) oder einen Benutzer (nach Anmeldedaten) darstellen.

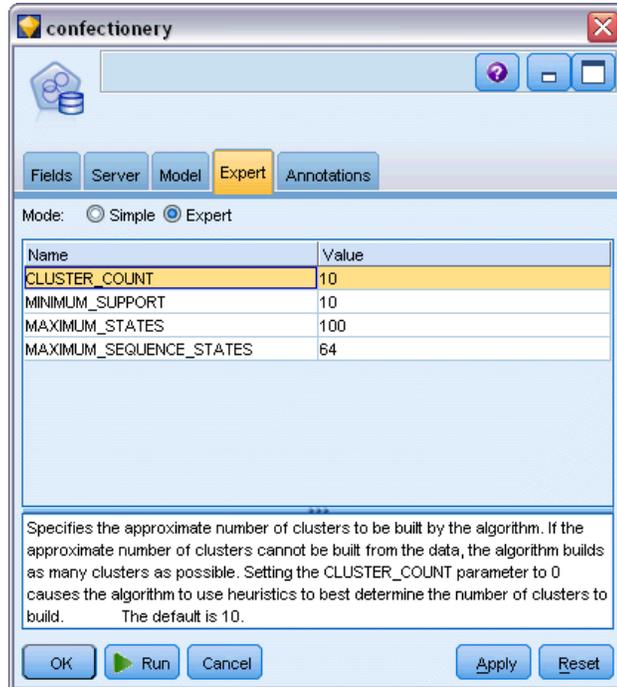
**Eingaben.** Wählen Sie das Eingabefeld bzw. die Eingabefelder für das Modell aus. Diese Felder enthalten die in der Sequenzmodellierung interessanten Ereignisse.

**Sequenz.** Wählen Sie ein Feld aus der Liste, das als Sequenz-ID-Feld verwendet werden soll. Sie können beispielsweise eine Webseiten-ID, eine Ganzzahl oder eine Zeichenkette verwenden, solange das Feld Ereignisse in einer Sequenz identifiziert. Nur eine Sequenz-ID ist pro Sequenz zulässig und nur ein Sequenztyp ist pro Modell erlaubt. Das Feld "Sequenz" muss sich vom Feld "ID" (in dieser Registerkarte) und vom Feld "Eindeutig" (in der Registerkarte "Modell") unterscheiden.

**Ziel.** Wählen Sie ein Feld, das als Zielfeld benutzt werden soll, d. h. das Feld, dessen Wert Sie auf der Grundlage der Sequenzdaten vorhersagen möchten.

### MS Sequenz-Clustering - Expertenoptionen

Abbildung 3-18  
Angabe der Expertenoptionen für MS Sequenz-Clustering



Die auf der Registerkarte "Experten" verfügbaren Optionen können je nach der Struktur des ausgewählten Streams variieren. Vollständige Details zu den für den ausgewählten Modellknoten von Analysis Server verfügbaren Expertenoptionen finden sie in der Hilfe zu den einzelnen Feldern in der Benutzeroberfläche.

### Scoring von Analysis Services-Modellen

Das Modell-Scoring erfolgt in SQL Server und wird durch Analysis Services durchgeführt. Wenn die Daten aus IBM® SPSS® Modeler stammen oder wenn diese in SPSS Modeler vorbereitet werden müssen, muss das Daten-Set gegebenenfalls in eine temporäre Tabelle geladen werden. Bei Modellen, die Sie mithilfe des In-Database Mining von SPSS Modeler aus erstellen, handelt es sich tatsächlich um Remote-Modelle, die auf dem entfernten Data-Mining- oder Datenbankserver gespeichert werden. Dies ist eine wichtige Unterscheidung, die beim Durchsuchen und Scoren von mit Microsoft Analysis Services-Algorithmen erstellten Modellen berücksichtigt werden muss.

In SPSS Modeler wird in der Regel nur eine Vorhersage mit der zugehörigen Wahrscheinlichkeit oder Konfidenz erstellt.

Weitere Beispiele zum Modell-Scoring finden Sie unter [Beispiele für das Mining mit Analysis Services auf S. 44](#).

## Gemeinsame Einstellungen für alle Analysis Services-Modelle

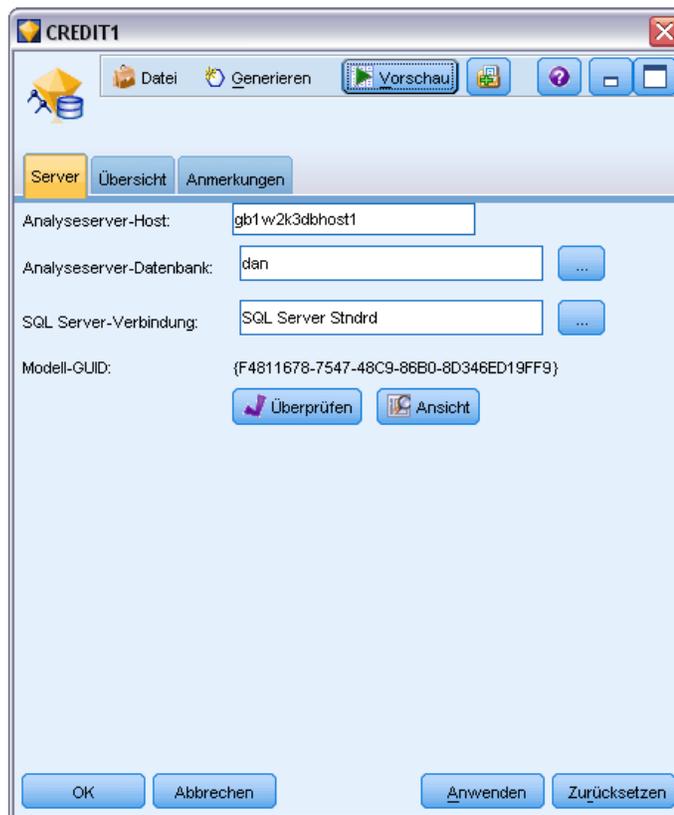
Folgende Einstellungen haben alle Analysis Services-Modelle gemeinsam.

### Analysis Services-Modell-Nugget – Registerkarte “Server”

Auf der Registerkarte “Server” können Verbindungen für das In-Database Mining angegeben werden. Auf der Registerkarte wird auch der eindeutige Modellschlüssel angegeben. Der Schlüssel wird bei der Modellerstellung nach dem Zufallsprinzip erzeugt und sowohl im Modell in IBM® SPSS® Modeler als auch in der Beschreibung des Modellobjekts gespeichert, das in der Analysis Services-Datenbank gespeichert ist.

Abbildung 3-19

Serveroptionen für den MS Entscheidungsbaum-Modell-Nugget

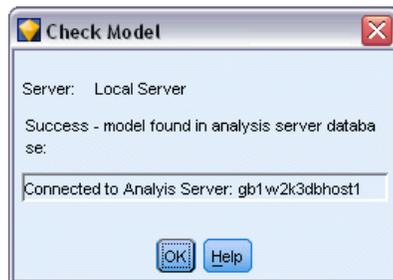


Auf der Registerkarte “Server” können Sie den Analyseserver-Host und die Analyseserver-Datenbank sowie die SQL Server-Datenquelle für den Scoring-Vorgang konfigurieren. Die hier festgelegten Optionen überschreiben die Optionen, die in den Dialogfeldern “Hilfsprogramme” oder “Modell erstellen” in IBM® SPSS® Modeler festgelegt wurden. Für weitere Informationen siehe [Thema Aktivieren der Integration mit Analysis Services](#) auf S. 15.

**Modell-GUID.** Hier wird der Modellschlüssel angezeigt. Der Schlüssel wird bei der Modellerstellung nach dem Zufallsprinzip erzeugt und sowohl im Modell in SPSS Modeler als auch in der Beschreibung des Modellobjekts gespeichert, das in der Analysis Services-Datenbank gespeichert ist.

**Überprüfen.** Klicken Sie auf diese Schaltfläche, um den Modellschlüssel mit dem Schlüssel des in der Analysis Services-Datenbank gespeicherten Modells zu vergleichen. Dadurch können Sie sicherstellen, dass das Modell noch im Analyseserver vorhanden ist und dass sich die Struktur des Modells nicht verändert hat.

Abbildung 3-20  
Ergebnisse der Überprüfung der Modellschlüssel

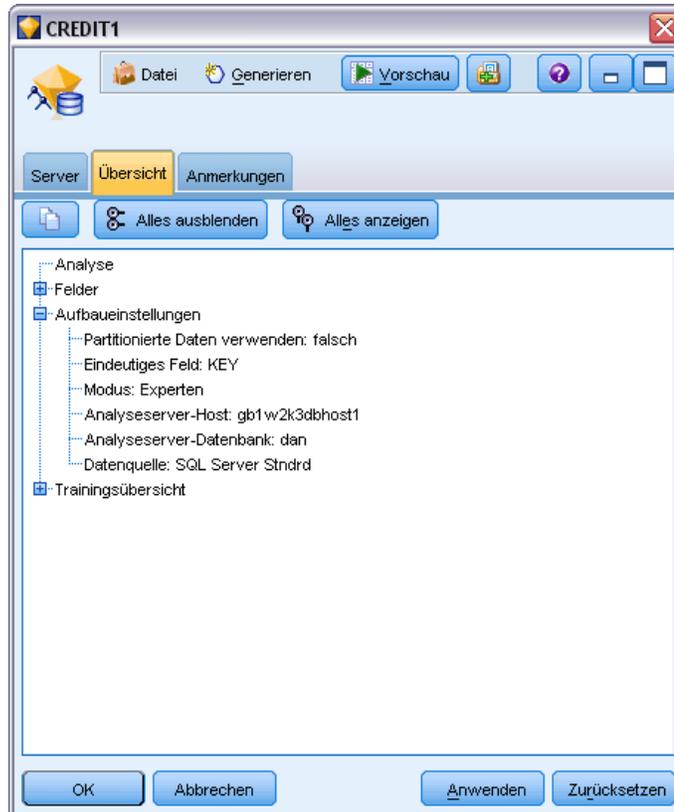


*Anmerkung:* Die Schaltfläche “Überprüfen” ist nur für Modelle verfügbar, die dem Stream zur Vorbereitung auf das Scoring hinzugefügt werden. Schlägt die Überprüfung fehl, stellen Sie fest, ob das Modell gelöscht oder durch ein anderes Modell auf dem Server ersetzt wurde.

**Ansicht.** Klicken Sie hier, um eine grafische Darstellung des Entscheidungsbaummodells zu erhalten. Der Entscheidungsbaum-Viewer steht auch für andere Entscheidungsbaumalgorithmen in SPSS Modeler zur Verfügung, wobei die Funktionalität immer gleich ist. [Für weitere Informationen siehe Thema Entscheidungsbaummodell-Nuggets in Kapitel 6 in IBM SPSS Modeler 14.2-Modellierungsknoten.](#)

### Analysis Services-Modell-Nugget – Registerkarte “Übersicht”

Abbildung 3-21  
Übersichtsoptionen für den MS Entscheidungsbaum-Modell-Nugget



Auf der Registerkarte “Übersicht” eines Modell-Nuggets finden Sie Informationen zum Modell selbst (*Analyse*), den im Modell verwendeten Feldern (*Felder*), den beim Erstellen des Modells verwendeten Einstellungen (*Aufbaueinstellungen*) und zum Modelltraining (*Trainingsübersicht*).

Beim ersten Durchsuchen des Knotens sind die Ergebnisse der Registerkarte “Übersicht” reduziert. Um die für Sie relevanten Ergebnisse anzuzeigen, vergrößern Sie sie mithilfe des Erweiterungssteuerelements links neben dem betreffenden Element oder klicken Sie auf die Schaltfläche *Alles anzeigen*, um alle Ergebnisse anzuzeigen. Um die Ergebnisse nach der Betrachtung wieder auszublenden, können Sie mit dem Erweiterungssteuerelement die gewünschten Ergebnisse reduzieren. Alternativ können Sie mit der Schaltfläche *Alles ausblenden* alle Ergebnisse ausblenden.

**Analyse.** Zeigt Informationen zum jeweiligen Modell an. Wenn Sie einen Analyseknoden ausgeführt haben, der an dieses Modell-Nugget angehängt ist, werden die Informationen aus dieser Analyse ebenfalls in diesem Abschnitt angezeigt. [Für weitere Informationen siehe Thema Analyseknoden in Kapitel 6 in IBM SPSS Modeler 14.2- Quellen-, Prozess- und Ausgabeknoten.](#)

**Felder.** Listet die Felder auf, die bei der Erstellung des Modells als Ziel bzw. als Eingaben verwendet werden.

**Aufbaueinstellungen.** Enthält Informationen zu den beim Aufbau des Modells verwendeten Einstellungen.

**Trainingsübersicht.** In diesem Abschnitt werden folgende Informationen angezeigt: der Typ des Modells, der für seine Erstellung verwendete Stream, der Benutzer, der ihn erstellt hat, der Erstellungszeitpunkt und die für den Aufbau des Modells benötigte Zeit.

### **MS Zeitreihen-Modell-Nugget**

Das MS Zeitreihenmodell erzeugt Werte nur für die vorhergesagten Zeitperioden, nicht für die historischen Daten.

Folgende Felder werden dem Modell hinzugefügt:

<b>Feldname</b>	<b>Beschreibung</b>
\$M-Feld	Vorhergesagter Wert von <i>Feld</i>
\$Var-Feld	Berechnete Varianz von <i>Feld</i>
\$Stdev-Feld	-Standardabweichung von <i>Feld</i>

### **MS Zeitreihen-Modell-Nugget – Registerkarte “Server”**

Auf der Registerkarte “Server” können Verbindungen für das In-Database Mining angegeben werden. Auf der Registerkarte wird auch der eindeutige Modellschlüssel angegeben. Der Schlüssel wird bei der Modellerstellung nach dem Zufallsprinzip erzeugt und sowohl im Modell in IBM® SPSS® Modeler als auch in der Beschreibung des Modellobjekts gespeichert, das in der Analysis Services-Datenbank gespeichert ist.

Abbildung 3-22  
Serveroptionen für das Modell-Nugget der MS Zeitreihen

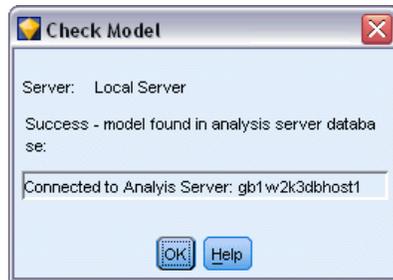


Auf der Registerkarte “Server” können Sie den Analyseserver-Host und die Analyseserver-Datenbank sowie die SQL Server-Datenquelle für den Scoring-Vorgang konfigurieren. Die hier festgelegten Optionen überschreiben die Optionen, die in den Dialogfeldern “Hilfsprogramme” oder “Modell erstellen” in IBM® SPSS® Modeler festgelegt wurden. [Für weitere Informationen siehe Thema Aktivieren der Integration mit Analysis Services auf S. 15.](#)

**Modell-GUID.** Hier wird der Modellschlüssel angezeigt. Der Schlüssel wird bei der Modellerstellung nach dem Zufallsprinzip erzeugt und sowohl im Modell in SPSS Modeler als auch in der Beschreibung des Modellobjekts gespeichert, das in der Analysis Services-Datenbank gespeichert ist.

**Überprüfen.** Klicken Sie auf diese Schaltfläche, um den Modellschlüssel mit dem Schlüssel des in der Analysis Services-Datenbank gespeicherten Modells zu vergleichen. Dadurch können Sie sicherstellen, dass das Modell noch im Analyseserver vorhanden ist und dass sich die Struktur des Modells nicht verändert hat.

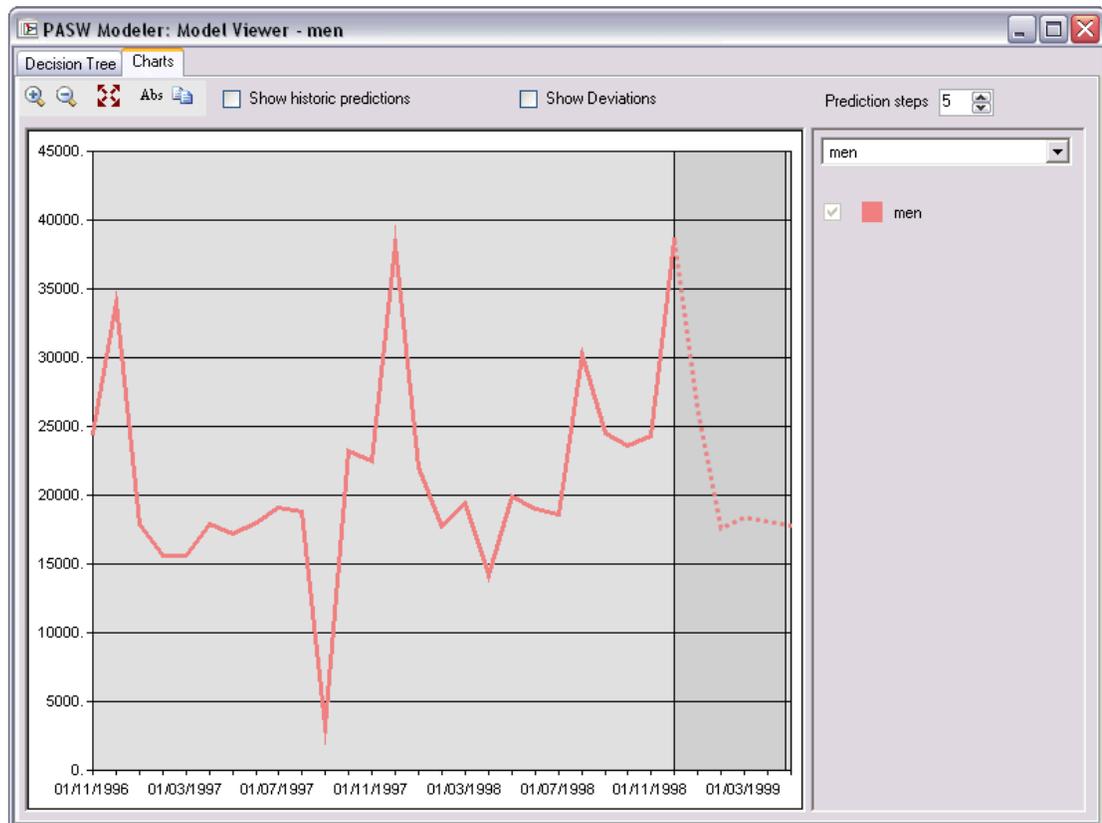
Abbildung 3-23.  
Ergebnisse der Überprüfung der Modellschlüssel



**Anmerkung:** Die Schaltfläche “Überprüfen” ist nur für Modelle verfügbar, die dem Stream zur Vorbereitung auf das Scoring hinzugefügt werden. Schlägt die Überprüfung fehl, stellen Sie fest, ob das Modell gelöscht oder durch ein anderes Modell auf dem Server ersetzt wurde.

**Ansicht.** Klicken Sie hier, um eine grafische Darstellung des Zeitreihenmodells zu erhalten. Analysis Services zeigt das fertiggestellte Modell als Baumstruktur an. Sie können auch ein Diagramm mit dem historischen Wert des Zielfelds im Laufe der Zeit zusammen mit vorhergesagten zukünftigen Werten anzeigen.

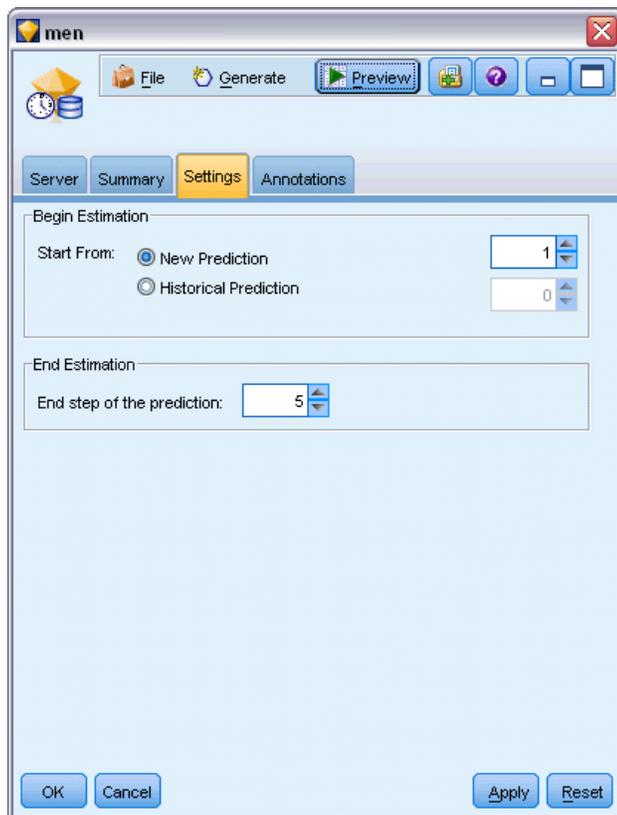
Abbildung 3-24  
MS Time Series-Viewer mit historischen (durchgehende Linie) und vorhergesagten zukünftigen Werten (gepunktete Linie)



Weitere Informationen finden Sie in der Beschreibung des Time Series-Viewer in der MSDN-Bibliothek unter <http://msdn.microsoft.com/en-us/library/ms175331.aspx>.

### MS Zeitreihen-Modell-Nugget – Registerkarte "Einstellungen"

Abbildung 3-25  
Einstellungsoptionen für das Modell-Nugget der MS Zeitreihen



**Schätzung beginnen.** Geben Sie die Zeitperiode an, in der Vorhersagen beginnen sollen.

- **Start: Neue Vorhersage.** Die Zeitperiode, in der zukünftige Vorhersagen beginnen sollen, ausgedrückt als Versatz von der letzten Zeitperiode Ihrer historischen Daten. Beispiel: Wenn Ihre historischen Daten bei 12.99 enden und Ihre Vorhersagen 01.00 beginnen sollen, verwenden Sie den Wert 1. Wenn die Vorhersagen jedoch 03.00 beginnen sollen, verwenden Sie den Wert 3.
- **Start: Historische Vorhersage.** Die Zeitperiode, in der historische Vorhersagen beginnen sollen, ausgedrückt als negativer Versatz von der letzten Zeitperiode Ihrer historischen Daten. Beispiel: Wenn Ihre historischen Daten bei 12.99 enden und Sie historische Vorhersagen für die letzten fünf Zeitperioden Ihrer Daten erstellen wollen, verwenden Sie den Wert -5.

**Schätzung beenden.** Geben Sie die Zeitperiode an, in der Vorhersagen enden sollen.

- **Letzter Schritt der Vorhersage.** Die Zeitperiode, in der zukünftige Vorhersagen enden sollen, ausgedrückt als Versatz von der letzten Zeitperiode Ihrer historischen Daten. Beispiel: Wenn Ihre historischen Daten bei 12.99 enden und Ihre Vorhersagen 6.00 enden sollen, verwenden

Sie hier den Wert 6. Für zukünftige Vorhersagen muss der Wert stets größer oder gleich dem Start-Wert sein.

### **MS Sequenz-Clustering-Modell-Nugget**

Die folgenden Felder werden dem MS Sequenz-Clustering-Modell hinzugefügt (dabei bezeichnet *Feld* den Namen des Zielfelds):

<b>Feldname</b>	<b>Beschreibung</b>
\$MC-Feld	Vorhersage des Clusters, dem diese Sequenz angehört.
\$MCP-Feld	Wahrscheinlichkeit, dass diese Sequenz dem vorhergesagten Cluster angehört.
\$MS-Feld	Vorhergesagter Wert von <i>Feld</i>
\$MSP-Feld	Wahrscheinlichkeit, dass der Wert von <i>\$MS-Feld</i> korrekt ist.

### **Exportieren von Modellen und Generieren von Knoten**

Sie können eine Modellübersicht und -struktur als Text- oder HTML-Datei exportieren. Sie können die benötigten Auswahl- und Filterknoten generieren. [Für weitere Informationen siehe Thema Durchsuchen von Modell-Nuggets in Kapitel 3 in IBM SPSS Modeler 14.2-Modellierungsknoten.](#)

Ähnlich wie andere Modell-Nuggets in IBM® SPSS® Modeler unterstützen die Microsoft Analysis Services-Modell-Nuggets die direkte Erzeugung von Datensatz- und Feldoperationsknoten. Mit den Optionen im Menü “Generieren” des Modell-Nuggets können Sie die folgenden Knoten erzeugen:

- Auswahlknoten (nur wenn auf der Registerkarte “Modell” ein Element ausgewählt ist)
- Filterknoten

### **Beispiele für das Mining mit Analysis Services**

Im Lieferumfang sind einige Beispiel-Streams enthalten, die die Verwendung von MS Analysis Services Data-Mining mit IBM® SPSS® Modeler demonstrieren. Diese Streams befinden sich im SPSS Modeler-Installationsordner unter:

`\Demos\Database_Modelling\Microsoft`

*Hinweis:* Dieser Demo-Ordner kann über die Programmgruppe “IBM SPSS Modeler” im Windows-Startmenü aufgerufen werden.

## Beispiel-Streams: Decision Trees (Entscheidungsbäume)

Die folgenden Streams können zusammen in Folge verwendet werden als Beispiel für einen Database-Mining-Prozess, bei dem der Entscheidungsbaumalgorithmus von MS Analysis Services verwendet wird.

Stream	Beschreibung
<i>1_upload_data.str</i>	Bereinigt Daten und lädt sie aus einer Textdatei in die Datenbank.
<i>2_explore_data.str</i>	Bietet ein Beispiel für die Datenuntersuchung mit IBM® SPSS® Modeler.
<i>3_build_model.str</i>	Erstellt das Modell unter Verwendung des datenbankeigenen Algorithmus.
<i>4_evaluate_model.str</i>	Wird als Beispiel für die Modellevaluation mit SPSS Modeler verwendet.
<i>5_deploy_model.str</i>	Verwendet das Modell für datenbankinternes Scoring.

*Anmerkung:* Um das Beispiel auszuführen, müssen die Streams in der richtigen Reihenfolge ausgeführt werden. Außerdem müssen die Quellen- und Modellierungsknoten in den einzelnen Streams aktualisiert werden, um auf eine gültige Datenquelle für die zu verwendende Datenbank zu verweisen.

Die in den Beispiel-Streams verwendeten Datenmengen beziehen sich auf Kreditkartenanwendungen und stellen ein Klassifizierungsproblem mit einer Mischung aus kategorialen und stetigen Prädiktoren dar. Weitere Informationen über diese Datenmenge finden Sie in der Datei *crx.names* im selben Ordner wie die Beispiel-Streams.

Diese Daten stehen im UCI Machine Learning Repository unter der folgenden Adresse zur Verfügung: <ftp://ftp.ics.uci.edu/pub/machine-learning-databases/credit-screening/>.

### Beispiel-Stream: Hochladen von Daten

Der erste Beispiel-Stream, *1\_upload\_data.str*, wird verwendet, um Daten aus einer Textdatei zu bereinigen und in SQL Server hochzuladen.

Abbildung 3-26  
Beispiel-Stream zum Hochladen von Daten



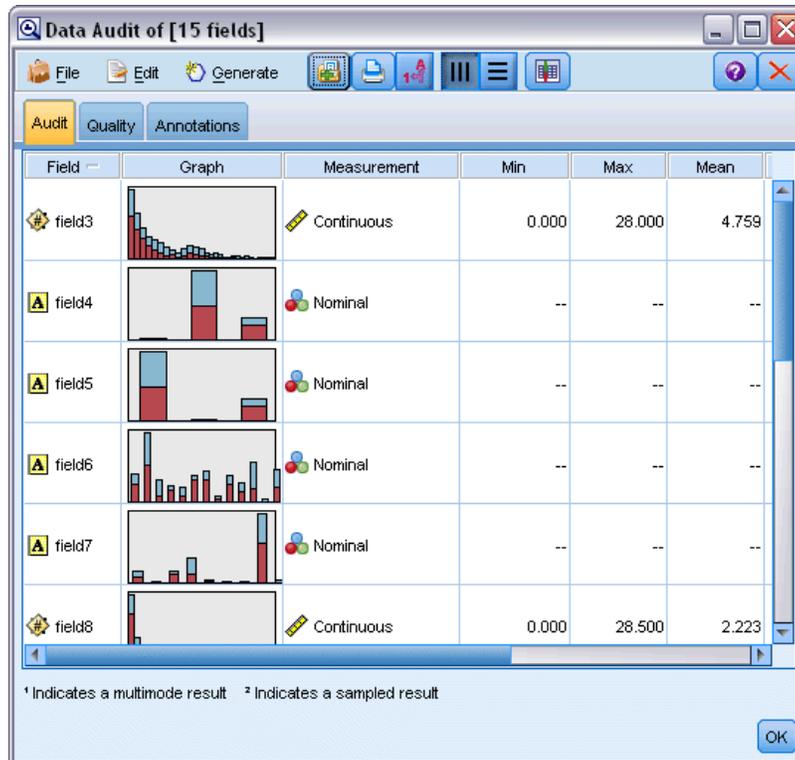
Da für das Data-Mining mit Analysis Services ein Schlüsselfeld erforderlich ist, fügt dieser erste Stream mithilfe eines Ableitungsknotens und der @INDEX-Funktion von IBM® SPSS® Modeler dem Daten-Set ein neues Feld mit dem Namen *KEY* und den eindeutigen Werten *1,2,3* hinzu.

Der nachfolgende Füllerknoten ist für die Behandlung von fehlenden Werten zuständig und ersetzt leere, aus der Textdatei *crx.data* eingelesene Felder durch *NULL*-Werte.

**Beispiel-Stream: Untersuchen von Daten**

Der zweite Beispiel-Stream, *2\_explore\_data.str*, soll zeigen, wie mithilfe eines Data Audit-Knotens ein allgemeiner Überblick über die Daten (einschließlich statistischer Funktionen und Diagramme) gewonnen werden kann. Für weitere Informationen siehe Thema [Data Audit-Knoten in Kapitel 6 in IBM SPSS Modeler 14.2- Quellen-, Prozess- und Ausgabeknoten](#).

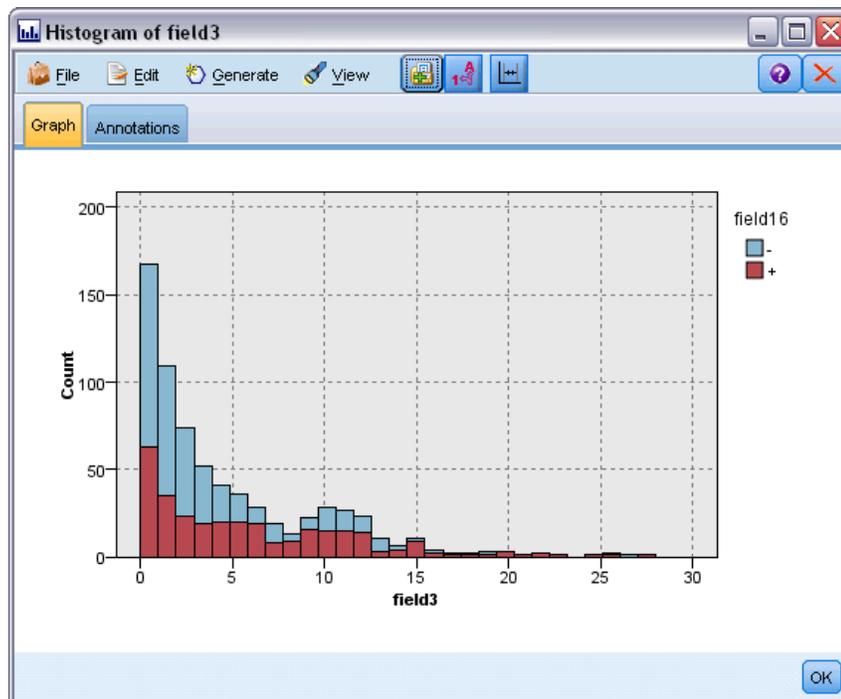
Abbildung 3-27  
Data Audit-Ergebnisse



Wenn Sie im Data Audit-Bericht auf ein Diagramm doppelklicken, wird ein detaillierteres Diagramm angezeigt, in dem Sie einzelne Felder eingehender untersuchen können.

Abbildung 3-28

Im Fenster "Data Audit" durch Doppelklicken auf dem Diagramm erzeugtes Histogramm

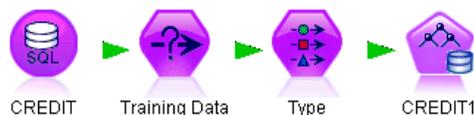


### Beispiel-Stream: Erstellen des Modells

Der dritte Beispiel-Stream, *3\_build\_model.str*, veranschaulicht die Modellerstellung in IBM® SPSS® Modeler. Sie können das Datenbankmodell an den Stream anhängen und darauf doppelklicken, um Einstellungen für die Erstellung festzulegen.

Abbildung 3-29

Beispiel-Stream für die Datenbank-Modellierung. Die violett eingezeichneten Knoten werden in der Datenbank ausgeführt.



Auf der Registerkarte "Modell" des Dialogfelds können Sie Folgendes festlegen:

- Wählen Sie das Feld Key als eindeutiges ID-Feld.

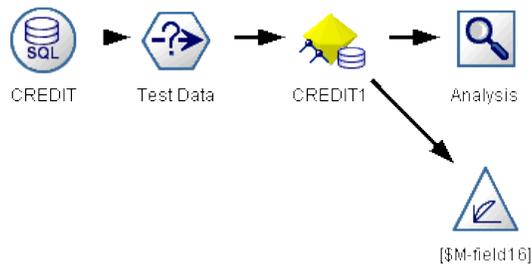
Auf der Registerkarte "Experten" können Sie die Einstellungen für die Modellerstellung verfeinern.

Stellen Sie vor dem Ausführen des Streams sicher, dass Sie die richtige Datenbank für die Modellerstellung angegeben haben. Verwenden Sie die Registerkarte "Server", um beliebige Einstellungen zu berichtigen.

### **Beispiel-Stream: Evaluieren des Modells**

Der vierte Beispiel-Stream, *4\_evaluate\_model.str*, veranschaulicht die Vorteile der Verwendung von IBM® SPSS® Modeler für die datenbankinterne Modellbildung. Sobald Sie das Modell ausgeführt haben, können Sie es wieder zu Ihrem Daten-Stream hinzufügen und das Modell mit verschiedenen von SPSS Modeler bereitgestellten Tools evaluieren.

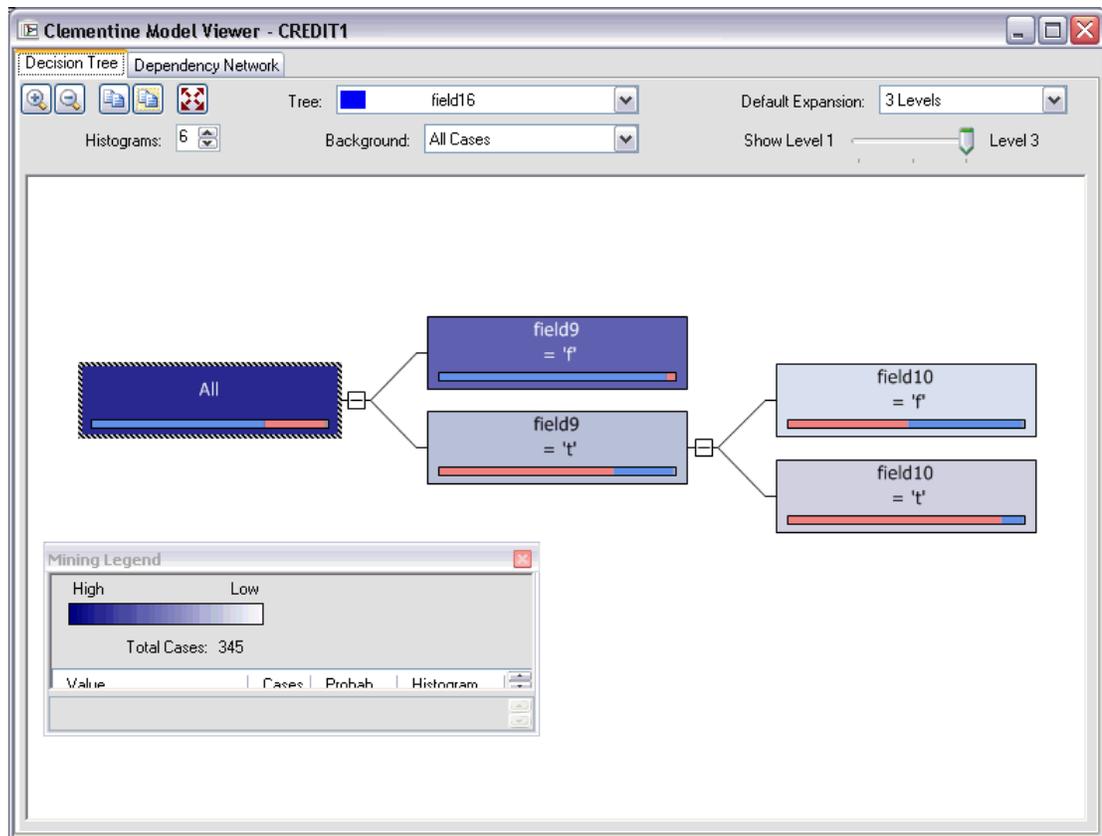
Abbildung 3-30  
Beispiel-Stream für die Modellevaluation



### **Anzeigen der Modellbildungsergebnisse**

Sie können durch einen Doppelklick auf das Modell-Nugget Ihre Ergebnisse untersuchen. Auf der Registerkarte “Übersicht” werden die Ergebnisse in einer Baumansicht angezeigt. Mit der Schaltfläche Ansicht auf der Registerkarte “Server” öffnen Sie eine grafische Darstellung des Entscheidungsbaummodells.

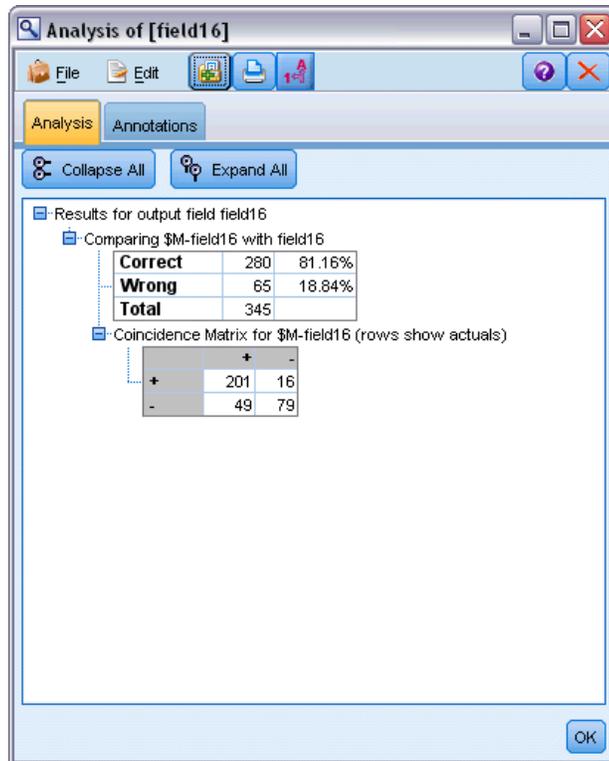
Abbildung 3-31  
Viewer mit grafischer Darstellung der MS Modellergebnisse aus den Entscheidungsbäumen



### **Evaluieren der Modellergebnisse**

Der Analyseknott im Beispiel-Stream erstellt eine Fehlklassifizierungstabelle, aus der das Muster der Übereinstimmungen zwischen jedem vorhergesagten Feld und dem zugehörigen Zielfeld ersichtlich wird. Führen Sie den Analyseknott aus, um die Ergebnisse anzuzeigen.

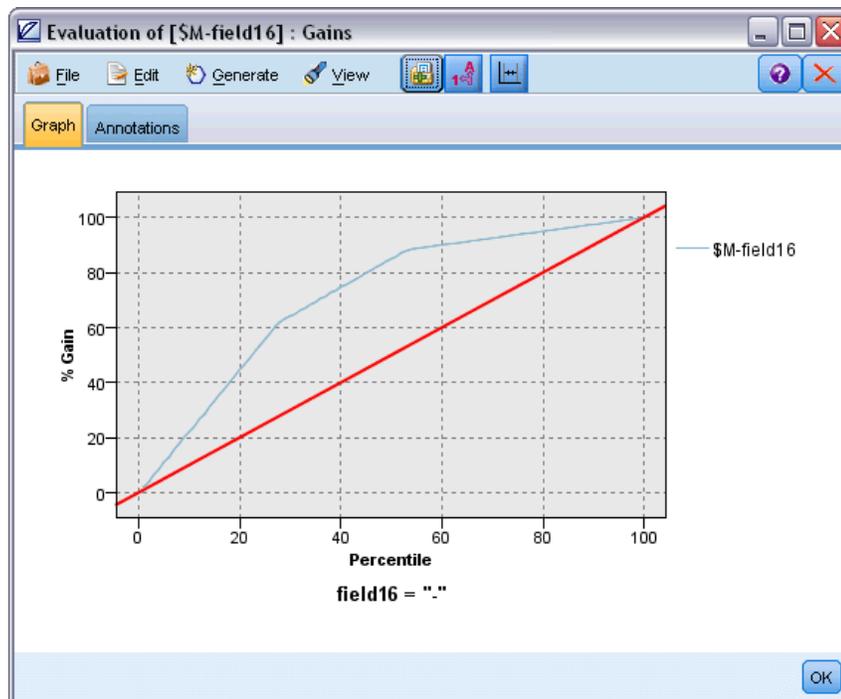
Abbildung 3-32  
Ergebnisse des Analyseknotens



Aus der Tabelle geht hervor, dass 81,16 % der vom MS Entscheidungsbaumalgorithmus generierten Vorhersagen richtig sind.

Der Evaluierungsknoten im Beispiel-Stream kann ein Gewinnendiagramm erstellen, das die Verbesserungen der Vorhersagegenauigkeit durch das Modell aufzeigt. Führen Sie den Evaluierungsknoten aus, um die Ergebnisse anzuzeigen.

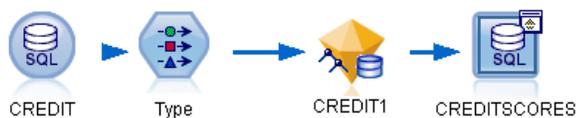
Abbildung 3-33  
Mit dem Evaluierungsknoten erstelltes Gewinn diagramm



### Beispiel-Stream: Bereitstellen des Modells

Sobald Sie mit der Genauigkeit des Modells zufrieden sind, können Sie es für die Verwendung mit externen Anwendungen oder für eine erneute Veröffentlichung in der Datenbank bereitstellen. Im letzten Beispiel-Stream, *5\_deploy\_model.str*, werden Daten aus der Tabelle CREDIT gelesen und dann mit dem Datenbankeexport-Knoten gescort und in der Tabelle CREDITSCORES veröffentlicht.

Abbildung 3-34  
Beispiel-Stream zum Bereitstellen des Modells



Durch Ausführen des Streams wird folgender SQL-Code erzeugt:

```
DROP TABLE CREDITSCORES
```

```
CREATE TABLE CREDITSCORES ( "field1" varchar(1),"field2" varchar(255),"field3" float,"field4" varchar(1),"field5" varchar(2),"field6" varchar(2),"field7" varchar(2),"field8" float,"field9" varchar(1),"field10" varchar(1),"field11" int,"field12" varchar(1),"field13" varchar(1),"field14" int,"field15" int,"field16" varchar(1),"KEY" int,"$M-field16" varchar(9),"$MC-field16" float )
```

```

INSERT INTO CREDITSCORES ("field1","field2","field3","field4","field5","field6","field7","field8",
"field9","field10","field11","field12","field13","field14","field15","field16",
"KEY","$M-field16","$MC-field16")
SELECT T0.C0 AS C0,T0.C1 AS C1,T0.C2 AS C2,T0.C3 AS C3,T0.C4 AS C4,T0.C5 AS C5,
    T0.C6 AS C6,T0.C7 AS C7,T0.C8 AS C8,T0.C9 AS C9,T0.C10 AS C10,
    T0.C11 AS C11,T0.C12 AS C12,T0.C13 AS C13,T0.C14 AS C14,
    T0.C15 AS C15,T0.C16 AS C16,T0.C17 AS C17,T0.C18 AS C18
FROM (
SELECT CONVERT(NVARCHAR,[TA].[field1]) AS C0, CONVERT(NVARCHAR,[TA].[field2]) AS C1,
    [TA].[field3] AS C2, CONVERT(NVARCHAR,[TA].[field4]) AS C3,
    CONVERT(NVARCHAR,[TA].[field5]) AS C4, CONVERT(NVARCHAR,[TA].[field6]) AS C5,
    CONVERT(NVARCHAR,[TA].[field7]) AS C6, [TA].[field8] AS C7,
    CONVERT(NVARCHAR,[TA].[field9]) AS C8, CONVERT(NVARCHAR,[TA].[field10]) AS C9,
    [TA].[field11] AS C10, CONVERT(NVARCHAR,[TA].[field12]) AS C11,
    CONVERT(NVARCHAR,[TA].[field13]) AS C12, [TA].[field14] AS C13,
    [TA].[field15] AS C14, CONVERT(NVARCHAR,[TA].[field16]) AS C15,
    [TA].[KEY] AS C16, CONVERT(NVARCHAR,[TA].[$M-field16]) AS C17,
    [TA].[$MC-field16] AS C18
FROM openrowset('MSOLAP',
'Datasource=localhost;Initial catalog=FoodMart 2000',
'SELECT [T].[C0] AS [field1],[T].[C1] AS [field2],[T].[C2] AS [field3],
    [T].[C3] AS [field4],[T].[C4] AS [field5],[T].[C5] AS [field6],
    [T].[C6] AS [field7],[T].[C7] AS [field8],[T].[C8] AS [field9],
    [T].[C9] AS [field10],[T].[C10] AS [field11],[T].[C11] AS [field12],
    [T].[C12] AS [field13],[T].[C13] AS [field14],[T].[C14] AS [field15],
    [T].[C15] AS [field16],[T].[C16] AS [KEY],[CREDIT1].[field16] AS [$M-field16],
    PredictProbability([CREDIT1].[field16]) AS [$MC-field16]
FROM [CREDIT1] PREDICTION JOIN
    openrowset('MSDASQL',
'Dsn=LocalServer;Uid=;pwd=','SELECT T0."field1" AS C0,T0."field2" AS C1,
    T0."field3" AS C2,T0."field4" AS C3,T0."field5" AS C4,T0."field6" AS C5,
    T0."field7" AS C6,T0."field8" AS C7,T0."field9" AS C8,T0."field10" AS C9,
    T0."field11" AS C10,T0."field12" AS C11,T0."field13" AS C12,
    T0."field14" AS C13,T0."field15" AS C14,T0."field16" AS C15,
    T0."KEY" AS C16 FROM "dbo".CREDITDATA T0') AS [T]
ON [T].[C2] = [CREDIT1].[field3] and [T].[C7] = [CREDIT1].[field8]
and [T].[C8] = [CREDIT1].[field9] and [T].[C9] = [CREDIT1].[field10]
and [T].[C10] = [CREDIT1].[field11] and [T].[C11] = [CREDIT1].[field12]
and [T].[C14] = [CREDIT1].[field15]') AS [TA]
) TO

```

# ***Datenbank-Modellbildung mit Oracle Data Mining***

## ***Informationen zu Oracle Data Mining***

IBM® SPSS® Modeler unterstützt die Integration mit Oracle Data Mining (ODM), das eine Serie von eng in Oracle RDBMS integrierten Data-Mining-Algorithmen bietet. Der Zugriff auf diese Funktionen erfolgt über die grafische Benutzeroberfläche und die am Workflow orientierte Entwicklungsumgebung von SPSS Modeler. So können Kunden die Data-Mining-Algorithmen von ODM verwenden.

SPSS Modeler unterstützt die Integration folgender Oracle Data Mining-Algorithmen:

- Naive Bayes
- Adaptive Bayes
- Support Vector Machine (SVM)
- Verallgemeinerte lineare Modelle (GLM)\*
- Entscheidungsbaum
- O-Cluster
- k-Means
- Nonnegative Matrix Factorization (NMF)
- A Priori
- Minimum Descriptor Length (MDL)
- Attribute Importance (AI)

\* 11nur g R1

## ***Voraussetzungen für die Integration mit Oracle***

Für die datenbankinterne Modellbildung mit Oracle Data Mining gelten die folgenden Voraussetzungen. Wenden Sie sich ggf. an Ihren Datenbankverwalter, um sicherzustellen, dass diese Bedingungen erfüllt sind.

- Ausführung von IBM® SPSS® Modeler im lokalen Modus oder im Rahmen einer IBM® SPSS® Modeler Server-Installation unter Windows oder UNIX.
- Oracle 10gR2 oder 11gR1 (10.2 Database oder höher) mit der Option für Oracle Data Mining.

*Anmerkung:* 10gR2 bietet Unterstützung für alle Datenbankmodellierungsalgorithmen außer “Verallgemeinerte lineare Modelle” (erfordert 11gR1).

- Eine ODBC-Datenquelle für die Verbindung mit Oracle, wie unten beschrieben.

*Hinweis:* Für Datenbankmodellierung und SQL-Optimierung muss auf dem SPSS Modeler-Computer die SPSS Modeler Server-Konnektivität aktiviert sein. Wenn diese Einstellung aktiviert ist, können Sie auf Datenbankalgorithmen zugreifen, SQL direkt aus SPSS Modeler per Pushback übertragen und auf SPSS Modeler Server zugreifen. Um den aktuellen Lizenzstatus zu überprüfen, klicken Sie im Menü “Hilfe” in SPSS Modeler auf Info. [Für weitere Informationen siehe Thema Verbindung mit IBM SPSS Modeler Server in Kapitel 3 in IBM SPSS Modeler 14.2- Benutzerhandbuch.](#)

## **Aktivieren der Integration mit Oracle**

Um die Integration von IBM® SPSS® Modeler mit Oracle Data Mining zu ermöglichen, müssen Sie Oracle konfigurieren, eine ODBC-Datenquelle erstellen, im SPSS Modeler-Dialogfeld “Hilfsprogramme” die Integration aktivieren und die SQL-Erzeugung und -Optimierung aktivieren.

### **Konfigurieren von Oracle**

Informationen zur Installation und Konfiguration von Oracle Data Mining finden Sie in der Oracle-Dokumentation – weitere Details enthält insbesondere der *Oracle Administrator's Guide*.

### **Erstellen einer ODBC-Datenquelle für Oracle**

Um die Verbindung zwischen Oracle und SPSS Modeler zu aktivieren, müssen Sie einen ODBC-Datenquellennamen (DSN) erstellen.

Bevor Sie einen DSN erstellen, sollten Sie grundlegende Kenntnisse über ODBC-Datenquellen und -Treiber sowie über Datenbankunterstützung in SPSS Modeler besitzen. [Für weitere Informationen siehe Thema Datenzugriff in Kapitel 2 in IBM SPSS Modeler Server 14.2-Verwaltungs- und Leistungshandbuch.](#)

Wenn Sie mit IBM® SPSS® Modeler Server im verteilten Modus arbeiten, müssen Sie den DSN auf dem Server-Computer erzeugen. Wenn Sie im lokalen (Client-)Modus arbeiten, müssen Sie auf dem Client-Computer einen DSN erzeugen.

- ▶ Installieren Sie die ODBC-Treiber. Diese Treiber finden Sie auf dem zu dieser Version gehörenden IBM® SPSS® Data Access Pack-Installationsmedium. Führen Sie die Datei *setup.exe* aus, um das Installationsprogramm zu starten und wählen Sie alle relevanten Treiber aus. Folgen Sie den Anweisungen am Bildschirm, um die Treiber zu installieren.
- ▶ Erstellen Sie den DSN.

*Anmerkung:* Die Befehlsfolge ist abhängig von der jeweiligen Windows-Version.

- **Windows XP.** Wählen Sie im Menü “Start” die Option Systemsteuerung. Doppelklicken Sie auf Verwaltung und dann auf Datenquellen (ODBC).

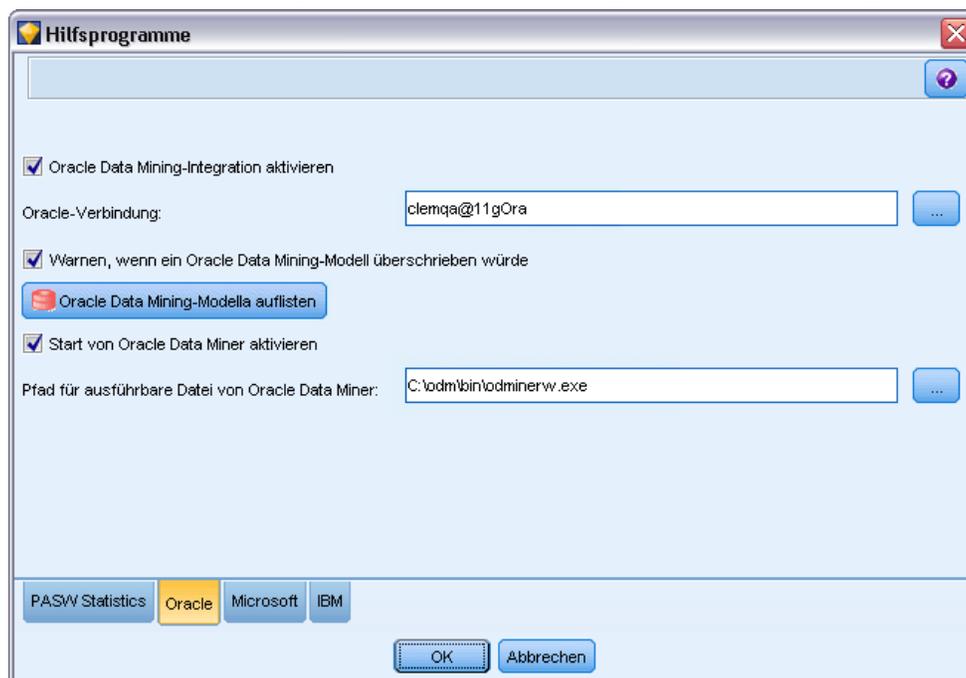
- **Windows Vista** Wählen Sie im Menü “Start” die Option Systemsteuerung und dann Systemwartung. Doppelklicken Sie auf Verwaltung, wählen Sie dann Datenquellen (ODBC) und klicken Sie auf Öffnen.
  - **Windows 7.** Wählen Sie im Menü “Start” die Option Systemsteuerung, dann System& Sicherheit und dann Verwaltung. Wählen Sie Datenquellen (ODBC) und klicken Sie dann auf Öffnen.
- ▶ Klicken Sie auf die Registerkarte System-DSN und dann auf Hinzufügen.
  - ▶ Wählen Sie den Treiber SPSS OEM 6.0 Oracle Wire Protocol aus.
  - ▶ Klicken Sie auf Fertig stellen.
  - ▶ Geben Sie im Bildschirm “ODBC Oracle Wire Protocol Driver Setup” einen Datenquellennamen Ihrer Wahl, den Hostnamen des Oracle-Servers, die Portnummer für die Verbindung und die SID für die verwendete Oracle-Instanz ein.
- Hostnamen, Port und SID finden Sie auf dem Server-Rechner in der Datei *tnsnames.ora*, sofern Sie TNS mit einer *tnsnames.ora*-Datei implementiert haben. Weitere Informationen erhalten Sie von Ihrem Oracle-Administrator.
- ▶ Klicken Sie auf die Schaltfläche Test, um die Verbindung zu testen.

#### **Aktivieren der Oracle Data Mining-Integration in IBM SPSS Modeler**

- ▶ Wählen Sie in den SPSS Modeler-Menüs folgende Befehlsfolge:  
Werkzeuge > Optionen > Hilfsprogramme
- ▶ Klicken Sie auf die Registerkarte Oracle.

Abbildung 4-1

Optionen für die Aktivierung der Datenbank-Modellbildung mit Oracle Data Mining



**Oracle Data Mining-Integration aktivieren.** Aktiviert die Datenbank-Modellbildungspalette (sofern diese nicht bereits angezeigt wird) und fügt Modellierungsknoten für Oracle Naive Bayes, Adaptive Bayes und Support Vector Machine hinzu. (Nach der Aktivierung wird diese Palette zusammen mit den anderen am unteren Rand des SPSS Modeler-Fensters angezeigt.)

**Oracle-Verbindung.** Legen Sie die Oracle ODBC-Datenquelle fest, die bei der Bildung und dem Speichern der Modelle als Standard benutzt wird, und geben Sie einen gültigen Benutzernamen und ein Passwort ein. Bei den einzelnen Modellierungsknoten und Modell-Nuggets kann diese Einstellung überschrieben werden.

*Hinweis:* Die für die Modellbildung benutzte Datenbankverbindung kann, muss aber nicht mit der für den Datenzugriff benutzten übereinstimmen. Sie können beispielsweise einen Stream einsetzen, der auf die Daten einer Oracle-Datenbank zugreift, die Daten für die Bereinigung und sonstige Bearbeitungen in SPSS Modeler herunterlädt und dann zur Modellbildung in eine andere Oracle-Datenbank lädt. Alternativ können sich die Originaldaten in einer Textdatei oder einer anderen Oracle-externen Quelle befinden. In diesem Fall müssen sie zur Modellbildung in Oracle geladen werden. In allen Fällen werden die Daten automatisch in eine temporäre Tabelle geladen, die in der für die Modellbildung genutzten Datenbank angelegt wird.

**Warnen, wenn ein Oracle Data Mining-Modell überschrieben würde.** Wählen Sie diese Option, um sicherzustellen, dass in der Datenbank gespeicherte Modelle nicht von SPSS Modeler überschrieben werden, ohne dass eine Warnung ausgegeben wird.

**Oracle Data Mining-Modelle auflisten.** Zeigt die verfügbaren Data-Mining-Modelle an.

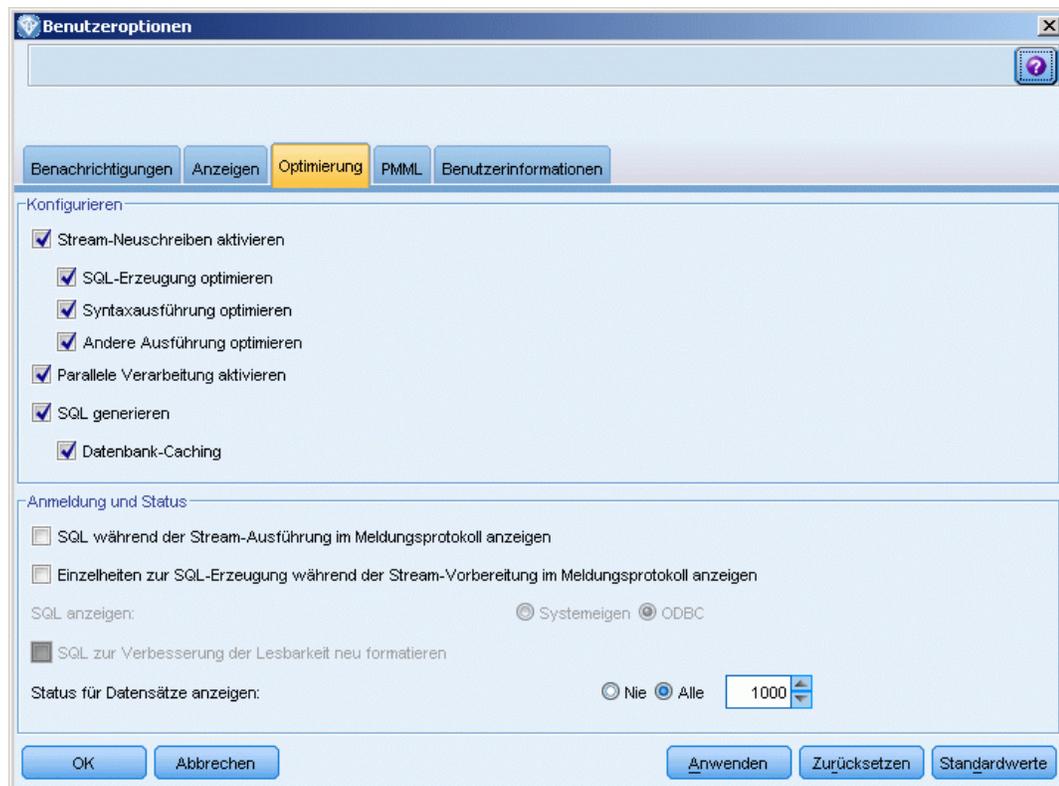
**Start von Oracle Data Miner aktivieren. (optional)** Wenn diese Option aktiviert ist, kann SPSS Modeler die Anwendung Oracle Data Miner starten. Weitere Informationen finden Sie unter [Oracle Data Miner auf S. 96](#).

**Pfad für ausführbare Datei von Oracle Data Miner. (optional)** Gibt den physischen Speicherort der ausführbaren Oracle Data Miner-Datei für Windows an (zum Beispiel `C:\odm\bin\odminerw.exe`). Oracle Data Miner wird nicht zusammen mit SPSS Modeler installiert. Die korrekte Version muss von der Oracle-Website (<http://www.oracle.com/technology/products/bi/odm/odminer.html>) heruntergeladen und auf dem Client installiert werden.

### ***Aktivieren der SQL-Erzeugung und -Optimierung***

- ▶ Wählen Sie in den SPSS Modeler-Menüs folgende Befehlsfolge:  
Werkzeuge > Optionen > Benutzeroptionen

Abbildung 4-2  
Optimierungseinstellungen



- ▶ Klicken Sie dann auf die Registerkarte Optimierung.
- ▶ Überzeugen Sie sich, dass die Option SQL generieren aktiviert ist. Diese Einstellung ist für die Datenbank-Modellierung erforderlich.
- ▶ Wählen Sie SQL-Erzeugung optimieren und Andere Ausführung optimieren aus. (Diese Einstellungen sind nicht unbedingt erforderlich, werden aber zur Leistungsoptimierung empfohlen.)

Für weitere Informationen siehe Thema Festlegen von Optimierungsoptionen in Kapitel 12 in *IBM SPSS Modeler 14.2- Benutzerhandbuch*.

## Modellbildung mit Oracle Data Mining

Die Oracle-Modellierungsknoten arbeiten bis auf einige wenige Ausnahmen in IBM® SPSS® Modeler genau wie andere Modellierungsknoten. Über die Datenbank-Modellbildungspalette am unteren Rand des SPSS Modeler-Fensters können Sie auf diese Knoten zugreifen.

Abbildung 4-3  
Datenbank-Modellbildungspalette



### **Erläuterung der Daten**

Für Oracle müssen kategoriale Daten in einem Zeichenkettenformat (entweder CHAR oder VARCHAR2) gespeichert sein. Demzufolge erlaubt SPSS Modeler nicht, dass numerische Speicherfelder mit dem Messniveau *Flag* oder *Nominal* (kategorial) als Eingabe für ODM-Modelle verwendet werden. Gegebenenfalls können Nummern in SPSS Modeler mit dem Umkodierungsknoten in Zeichenketten konvertiert werden. [Für weitere Informationen siehe Thema Umkodierungsknoten in Kapitel 4 in IBM SPSS Modeler 14.2- Quellen-, Prozess- und Ausgabeknoten.](#)

**Zielfeld.** In ODM-Klassifizierungsmodellen kann nur ein Feld als Ausgabefeld (Ziel) ausgewählt werden.

**Eindeutiges Feld.** Gibt das Feld an, über das jeder Fall eindeutig identifiziert wird. Dies kann beispielsweise ein ID-Feld sein, wie *CustomerID*. SPSS Modeler schreibt vor, dass dieses Schlüsselfeld zwingend numerisch sein muss.

*Anmerkung:* Dieses Feld ist für alle Oracle-Knoten mit Ausnahme von Oracle Adaptive Bayes, Oracle O-Cluster und Oracle Apriori optional.

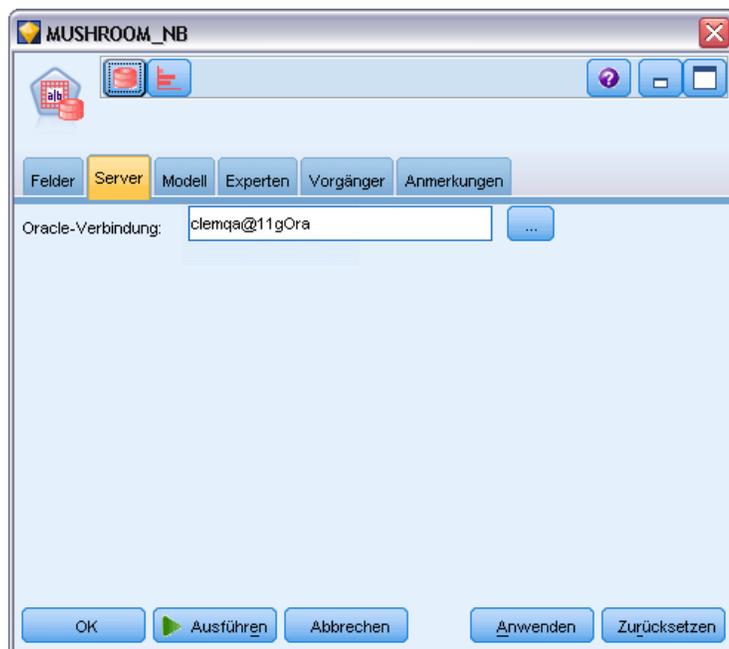
### **Allgemeine Kommentare**

- Für von Oracle Data Mining erstellte Modelle bietet SPSS Modeler keinen PMML-Export/-Import.
- Das Modell-Scoring erfolgt immer innerhalb von ODM. Wenn die Daten aus SPSS Modeler stammen oder dort vorbereitet werden müssen, muss das Daten-Set gegebenenfalls in eine temporäre Tabelle geladen werden.
- In SPSS Modeler wird in der Regel nur eine Vorhersage mit der zugehörigen Wahrscheinlichkeit oder Konfidenz erstellt.
- SPSS Modeler beschränkt die Anzahl der Felder, die beim Erstellen und Scoren von Modellen verwendet werden können, auf 1.000.
- SPSS Modeler kann ODM-Modelle mit IBM® SPSS® Modeler Solution Publisher aus zur Ausführung veröffentlichten Streams heraus scoren. [Für weitere Informationen siehe Thema So funktioniert IBM SPSS Modeler Solution Publisher in Kapitel 2 in IBM SPSS Modeler 14.2 Solution Publisher.](#)

## Serveroptionen für Oracle-Modelle

Legen Sie die Oracle-Verbindung fest, die zum Hochladen der für die Modellbildung verwendeten Daten benutzt wird. Gegebenenfalls können Sie auf der Registerkarte “Server” für jeden Modellierungsknoten eine Verbindung auswählen, mit der die im Dialogfeld “Hilfsprogramme” angegebene Standard-Oracle-Verbindung überschrieben wird. [Für weitere Informationen siehe Thema Aktivieren der Integration mit Oracle auf S. 54.](#)

Abbildung 4-4  
Oracle-Serveroptionen

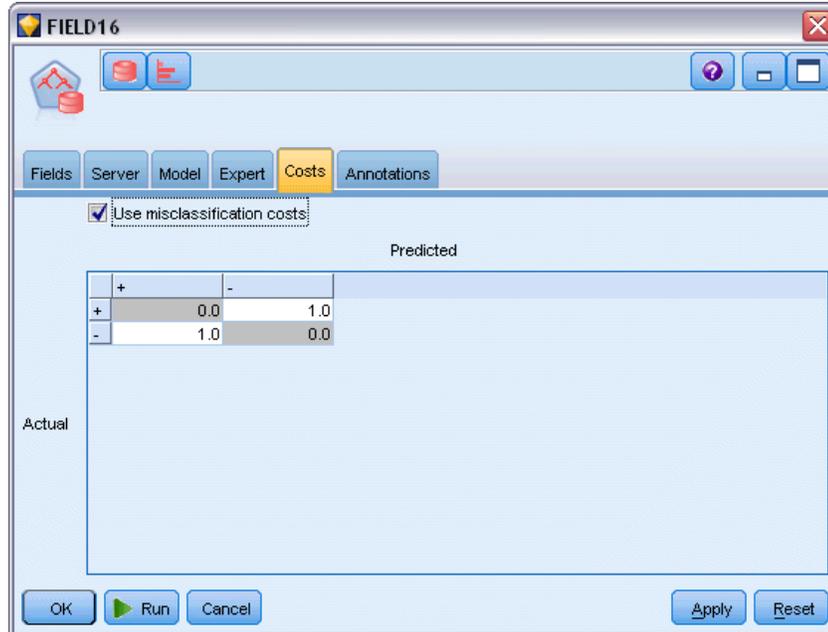


### Kommentare

- Die für die Modellierung benutzte Verbindung kann mit der im Quellenknoten für einen Stream benutzten Verbindung identisch sein. Sie können beispielsweise einen Stream einsetzen, der auf die Daten einer Oracle-Datenbank zugreift, die Daten für die Bereinigung und sonstige Bearbeitungen in IBM® SPSS® Modeler herunterlädt und dann zur Modellbildung in eine andere Oracle-Datenbank lädt.
- Der Name der ODBC-Datenquelle wird in jeden SPSS Modeler-Stream eingebettet. Wenn ein auf einem Host erzeugter Stream auf einem anderen Host ausgeführt wird, muss der Name der Datenquelle auf beiden Hosts identisch sein. Alternativ kann für jeden Quellen- oder Modellierungsknoten auf der Registerkarte “Server” eine andere Datenquelle ausgewählt werden.

## Fehlklassifizierungskosten

Abbildung 4-5  
Oracle-Kostenoptionen



In einigen Zusammenhängen sind bestimmte Fehler kostspieliger als andere. Zum Beispiel kann es kostspieliger sein, einen Kreditantragsteller mit hohem Risiko als niedriges Risiko zu klassifizieren (eine Art von Fehler) als einen Kreditantragsteller mit niedrigem Risiko als hohes Risiko (eine andere Art von Fehler). Anhand von Fehlklassifizierungskosten können Sie die relative Bedeutung verschiedener Arten von Vorhersagefehlern angeben.

Fehlklassifizierungskosten sind im Grunde Gewichtungen, die auf bestimmte Ergebnisse angewendet werden. Diese Gewichtungen werden in das Modell mit einbezogen und können die Vorhersage (als Schutz gegen kostspielige Fehler) ändern.

Mit Ausnahme von C5.0-Modellen werden Fehlklassifizierungskosten beim Scoring von Modellen nicht angewendet und bei der Rangbildung bzw. beim Vergleich von Modellen mithilfe eines Knotens vom Typ "Automatischer Klassifizierer", eines Evaluationsdiagramms oder eines Analyseknosens nicht berücksichtigt. Ein Modell, das Kosten beinhaltet, führt nicht unbedingt zu weniger Fehlern als ein Modell, das keine Kosten beinhaltet, und es weist auch nicht unbedingt eine höhere Gesamtgenauigkeit auf, aber es ist möglicherweise in praktischer Hinsicht leistungsfähiger, da es eine integrierte Verzerrung zugunsten von *weniger teuren* Fehlern aufweist.

Die Kostenmatrix zeigt die Kosten für jede mögliche Kombination aus prognostizierter Kategorie und tatsächlicher Kategorie. Standardmäßig werden alle Fehlklassifizierungskosten auf 1,0 gesetzt. Um benutzerdefinierte Kostenwerte einzugeben, wählen Sie Fehlklassifizierungskosten verwenden und geben Ihre benutzerdefinierten Werte in die Kostenmatrix ein.

Um die Fehlklassifizierungskosten zu ändern, wählen Sie die Zelle aus, die der gewünschten Kombination aus vorhergesagten und tatsächlichen Werten entspricht, löschen den Inhalt der Zelle und geben die gewünschten Kosten für die Zelle ein. Die Kosten sind nicht automatisch symmetrisch. Wenn Sie z. B. die Kosten einer Fehlklassifizierung von *A* als *B* auf 2,0 setzen,

weisen die Kosten für eine Fehlklassifizierung von  $B$  als  $A$  weiterhin den Standardwert 1,0 auf, es sei denn, Sie ändern diesen Wert ausdrücklich.

*Hinweis:* Nur im Entscheidungsbaummodell können die Kosten zum Zeitpunkt der Erstellung angegeben werden.

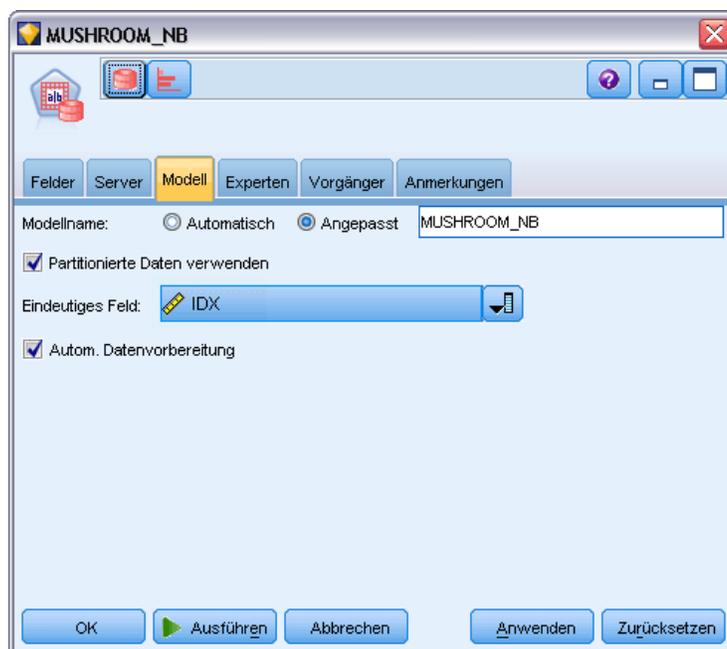
## Oracle Naive Bayes

Naive Bayes ist ein bekannter Algorithmus für Klassifizierungsprobleme. Das Modell wird als *naiv* bezeichnet, weil es alle vorgeschlagenen Vorhersagevariablen als voneinander unabhängig behandelt. Naive Bayes ist ein schneller, skalierbarer Algorithmus, der für Kombinationen von Attributen und für das Zielattribut konditionale Wahrscheinlichkeiten berechnet. Aus den Trainingsdaten wird eine unabhängige Wahrscheinlichkeit ermittelt. Diese liefert die Wahrscheinlichkeit jeder Zielklasse anhand des Vorkommens der einzelnen Wertekategorien aus jeder einzelnen Eingabevariablen.

- Die Kreuzvalidierung wird eingesetzt, um die Modellgenauigkeit mit denselben Daten zu testen, die zur Modellbildung verwendet wurden. Dies ist insbesondere dann nützlich, wenn die Anzahl der für die Modellbildung verfügbaren Fälle gering ist.
- Die Modellausgabe kann in einem Matrixformat durchsucht werden. Bei den in der Matrix vorhandenen Zahlen handelt es sich um bedingte Wahrscheinlichkeiten, die sich auf die vorhergesagten Fälle (Spalten) und die Variablen-Wert-Kombinationen der Prädiktoren (Zeilen) beziehen.

## Optionen für Naive Bayes-Modelle

Abbildung 4-6  
Optionen für Naive Bayes-Modelle



**Modellname.** Sie können den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen (oder dem Modelltyp in Fällen, in denen kein solches Feld angegeben ist) generieren oder einen benutzerdefinierten Namen eingeben.

**Partitionierte Daten verwenden.** Wenn ein Partitionsfeld definiert ist, gewährleistet diese Option, dass nur Daten aus der Trainingspartition für die Modellerstellung verwendet werden. [Für weitere Informationen siehe Thema Partitionsknoten in Kapitel 4 in IBM SPSS Modeler 14.2- Quellen-, Prozess- und Ausgabe-knoten.](#)

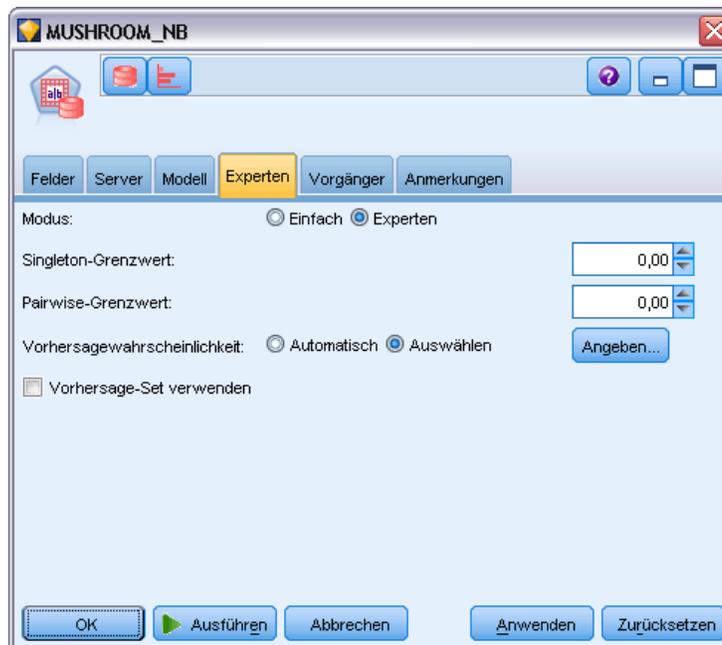
**Eindeutiges Feld.** Gibt das Feld an, über das jeder Fall eindeutig identifiziert wird. Dies kann beispielsweise ein ID-Feld sein, wie *CustomerID*. IBM® SPSS® Modeler schreibt vor, dass dieses Schlüsselfeld zwingend numerisch sein muss.

*Anmerkung:* Dieses Feld ist für alle Oracle-Knoten mit Ausnahme von Oracle Adaptive Bayes, Oracle O-Cluster und Oracle Apriori optional.

**Automatische Datenvorbereitung.** (Nur 11g only) Aktiviert (Standard) oder deaktiviert den automatisierten Datenvorbereitungsmodus von Oracle Data Mining. Wenn dieses Kästchen markiert ist, führt ODM automatisch die vom Algorithmus geforderten Datenumformungen durch. Weitere Informationen finden Sie unter *Oracle Data Mining Concepts*.

## Expertenoptionen für Naive Bayes

Abbildung 4-7  
Expertenoptionen für Naive Bayes



Wenn das Modell erstellt wird, werden einzelne Prädiktorattributwerte oder -wertpaare ignoriert, wenn ein bestimmter Wert oder ein Wertpaar in den Trainingsdaten nicht häufig genug vorkommt. Die Grenzwerte für ignorierte Werte werden als Anteilswerte der Anzahl der in den Trainingsdaten

vorhandenen Datensätze angeben. Die Anpassung dieser Grenzwerte kann das Rauschen reduzieren und die Voraussetzungen des Modells erhöhen, dass es auf andere Daten-Sets generalisiert werden kann.

- **Singleton-Grenzwert.** Legt den Grenzwert für einen bestimmten Prädiktorattributwert fest. Die Häufigkeit des Vorkommens eines bestimmten Werts muss gleich oder höher sein als der angegebene Anteilwert. Ansonsten wird der Wert ignoriert.
- **Pairwise-Grenzwert.** Legt den Grenzwert für ein bestimmtes Attribut und ein Prädiktorwertpaar fest. Die Häufigkeit des Vorkommens eines bestimmten Wertpaars muss gleich oder höher sein als der angegebene Anteilwert. Ansonsten wird das Paar ignoriert.

**Vorhersagewahrscheinlichkeit.** Ermöglicht dem Modell, die Wahrscheinlichkeit einer korrekten Prognose für ein mögliches Ergebnis des Zielfelds zu umfassen. Sie aktivieren diese Option, indem Sie Auswählen wählen, auf die Schaltfläche Angeben klicken, eines der möglichen Ergebnisse wählen und auf Einfügen klicken.

**Vorhersage-Set verwenden.** Generiert eine Tabelle von allen möglichen Resultaten für alle möglichen Ergebnisse des Zielfelds.

## Oracle Adaptive Bayes

Adaptive Bayes Network (ABN) bildet anhand der Minimum Description Length (MDL) und der automatischen Funktionsauswahl Bayesian Network Classifier. ABN funktioniert in bestimmten Situationen gut, in denen Naive Bayes wenig erreicht. In den meisten anderen Fällen funktioniert ABN mindestens genauso gut, die Leistung kann allerdings etwas langsamer sein. Mit dem ABN-Algorithmus können Baumtypen von erweiterten, auf Bayes basierenden Modellen gebildet werden, zu denen vereinfachte Entscheidungsbaummodelle (Einzelfunktion), reduzierte Naive Bayes-Modelle und verstärkte Multifunktionsmodelle gehören.

### Generierte Modelle

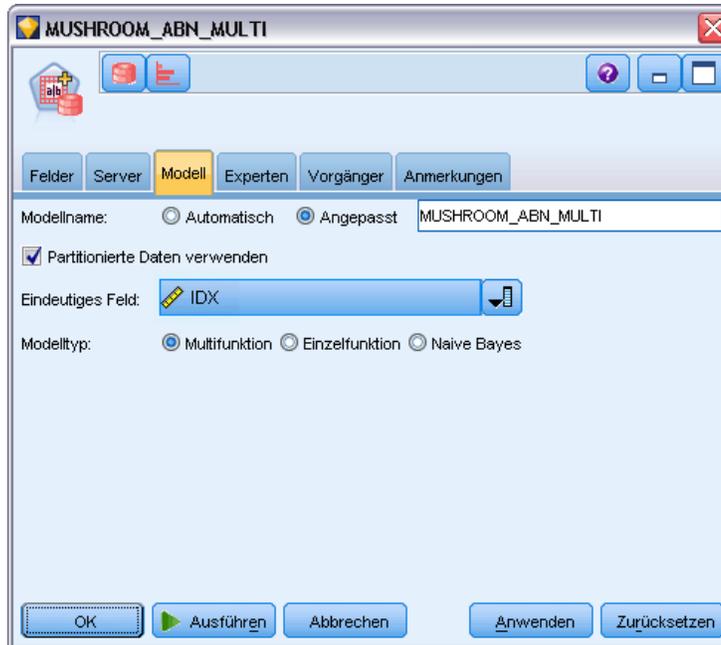
Im Einzelfunktionsmodus erzeugt ABN einen vereinfachten Entscheidungsbaum, der auf einem Set lesbarer Regeln basiert, über die Benutzer oder Analysten die Grundlage der Vorhersagen des Modells nachvollziehen und anderen entsprechend erläutern können. Dies kann sich im Vergleich zu Naive Bayes- oder Multifunktionsmodellen als signifikanter Vorteil erweisen. Diese Regeln können wie ein Standardregel-Set in IBM® SPSS® Modeler durchsucht werden. Ein einfaches Regel-Set könnte folgendermaßen aussehen:

```
IF MARITAL_STATUS = "Married"
AND EDUCATION_NUM = "13-16"
THEN CHURN= "TRUE"
Confidence = .78, Support = 570 cases
```

Reduzierte Naive Bayes- und Multifunktionsmodelle können nicht in SPSS Modeler durchsucht werden.

## Optionen für Adaptive Bayes-Modelle

Abbildung 4-8  
Optionen für Adaptive Bayes-Modelle



**Modellname.** Sie können den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen (oder dem Modelltyp in Fällen, in denen kein solches Feld angegeben ist) generieren oder einen benutzerdefinierten Namen eingeben.

**Partitionierte Daten verwenden.** Wenn ein Partitionsfeld definiert ist, gewährleistet diese Option, dass nur Daten aus der Trainingspartition für die Modellerstellung verwendet werden. [Für weitere Informationen siehe Thema Partitionsknoten in Kapitel 4 in IBM SPSS Modeler 14.2- Quellen-, Prozess- und Ausgabeknoten.](#)

**Eindeutiges Feld.** Gibt das Feld an, über das jeder Fall eindeutig identifiziert wird. Dies kann beispielsweise ein ID-Feld sein, wie *CustomerID*. IBM® SPSS® Modeler schreibt vor, dass dieses Schlüsselfeld zwingend numerisch sein muss.

*Anmerkung:* Dieses Feld ist für alle Oracle-Knoten mit Ausnahme von Oracle Adaptive Bayes, Oracle O-Cluster und Oracle Apriori optional.

### Modelltyp

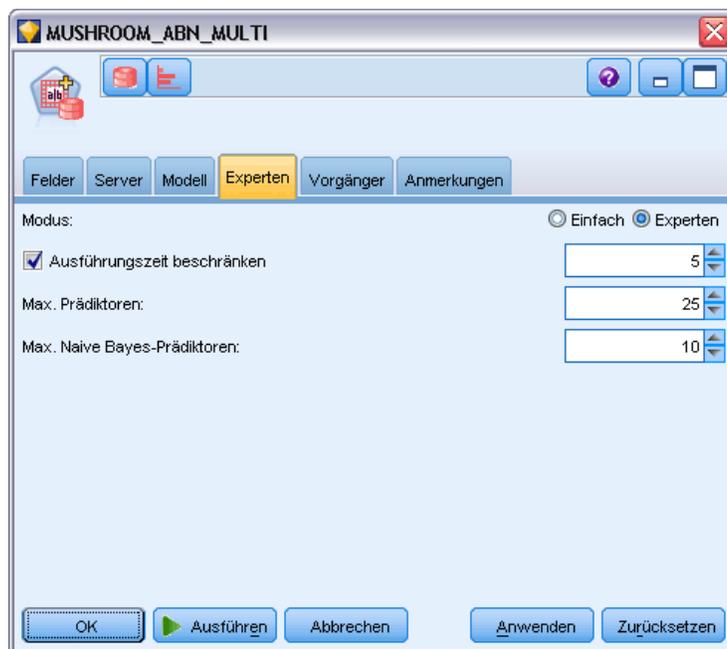
Zum Erstellen des Modells stehen drei verschiedene Modi zur Auswahl.

- **Multifunktion.** Erzeugt und vergleicht mehrere Modelle, einschließlich eines NB-Modells sowie Einzel- und Multifunktionsmodellen für die Produktwahrscheinlichkeit. Hierbei handelt es sich um den umfassendsten Modus, dessen Berechnung demnach in der Regel auch am längsten dauert. Regeln werden nur dann erzeugt, wenn sich das Einzelfunktionsmodell am besten eignet. Wenn ein Multifunktions- oder ein NB-Modell ausgewählt wird, werden keine Regeln erzeugt.

- **Einzelfunktion.** Erzeugt einen vereinfachten Entscheidungsbaum, der auf einem Set von Regeln basiert. Jede Regel enthält eine Bedingung und Wahrscheinlichkeiten, die jedem Ergebnis zugeordnet sind. Die Regeln sind untereinander exklusiv und liegen in einer für Menschen lesbaren Form vor, was sich gegenüber Naive Bayes- und Multifunktionsmodellen als signifikanter Vorteil erweisen kann.
- **Naive Bayes.** Erzeugt ein einzelnes NB-Modell und vergleicht dieses mit dem globalen Stichprobenvorgänger (die Verteilung der Zielwerte in der globalen Stichprobe). Das NB-Modell wird nur dann ausgegeben, wenn es sich für die Zielwerte als besserer Prädiktor erweist als der globale Vorgänger. Andernfalls wird das Modell nicht ausgegeben.

## Expertenoptionen für Adaptive Bayes

Abbildung 4-9  
Expertenoptionen für Adaptive Bayes



**Ausführungszeit beschränken.** Über diese Option können Sie eine maximale Erstellungszeit in Minuten angeben. Damit können Sie Modelle in kürzerer Zeit erstellen, wenngleich das daraus resultierende Modell weniger genau sein kann. An jeder Etappe des Modellbildungsverfahrens prüft der Algorithmus, bevor er fortfährt, ob er in der Lage ist, die nächste Etappe innerhalb der vorgegebenen Zeit abzuschließen, und liefert bei Erreichen der Zeitgrenze das beste verfügbare Modell zurück.

**Max. Prädiktoren.** Mit dieser Option können Sie die Komplexität des Modells einschränken und die Leistung verbessern, indem Sie die Anzahl der verwendeten Prädiktoren beschränken. Prädiktoren werden auf der Grundlage einer MDL-Messung ihrer Korrelation mit dem Ziel eingestuft. Diese Einstufung bestimmt die Wahrscheinlichkeit, dass sie in das Modell aufgenommen werden.

**Max. Naive Bayes-Prädiktoren.** Diese Option legt die maximale Anzahl der Prädiktoren fest, die im Naive Bayes-Modell verwendet werden.

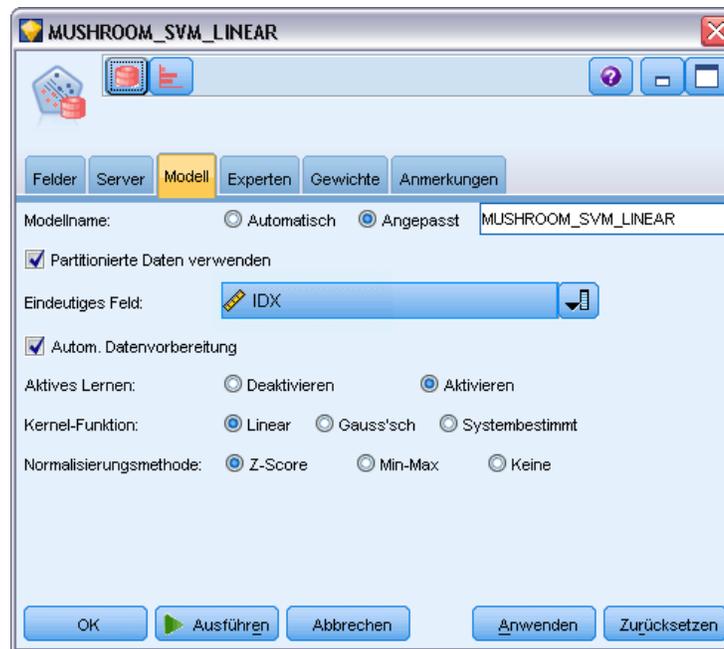
## Oracle Support Vector Machine (SVM)

Support Vector Machine (SVM) ist ein Klassifikations- und Regressionsalgorithmus, der eine Theorie des maschinellen Lernens nutzt, um die Vorhersagegenauigkeit zu maximieren, ohne die Daten übermäßig anzupassen. SVM nutzt eine optionale nichtlineare Transformation der Trainingsdaten, an die sich die Suche nach Regressionsgleichungen in den transformierten Daten anschließt, mit denen die Klassen getrennt werden (für kategoriale Ziele) oder das Ziel angepasst wird (für stetige Ziele). Die Oracle-Implementierung von SVM ermöglicht das Erstellen von Modellen unter Verwendung eines der zwei verfügbaren Kernels – des linearen oder des Gauss'schen Kernels. Der lineare Kernel verzichtet auf die gesamte nichtlineare Transformation und liefert als Modellergebnis im Grunde ein Regressionsmodell.

Weitere Informationen finden Sie in folgenden Publikationen: *Oracle Data Mining Application Developer's Guide* und *Oracle Data Mining Concepts*.

### Optionen für Oracle SVM-Modelle

Abbildung 4-10  
SVM-Modelloptionen



**Modellname.** Sie können den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen (oder dem Modelltyp in Fällen, in denen kein solches Feld angegeben ist) generieren oder einen benutzerdefinierten Namen eingeben.

**Eindeutiges Feld.** Gibt das Feld an, über das jeder Fall eindeutig identifiziert wird. Dies kann beispielsweise ein ID-Feld sein, wie *CustomerID*. IBM® SPSS® Modeler schreibt vor, dass dieses Schlüsselfeld zwingend numerisch sein muss.

*Anmerkung:* Dieses Feld ist für alle Oracle-Knoten mit Ausnahme von Oracle Adaptive Bayes, Oracle O-Cluster und Oracle Apriori optional.

**Automatische Datenvorbereitung.** (Nur 11g only) Aktiviert (Standard) oder deaktiviert den automatisierten Datenvorbereitungsmodus von Oracle Data Mining. Wenn dieses Kästchen markiert ist, führt ODM automatisch die vom Algorithmus geforderten Datenumformungen durch. Weitere Informationen finden Sie unter *Oracle Data Mining Concepts*.

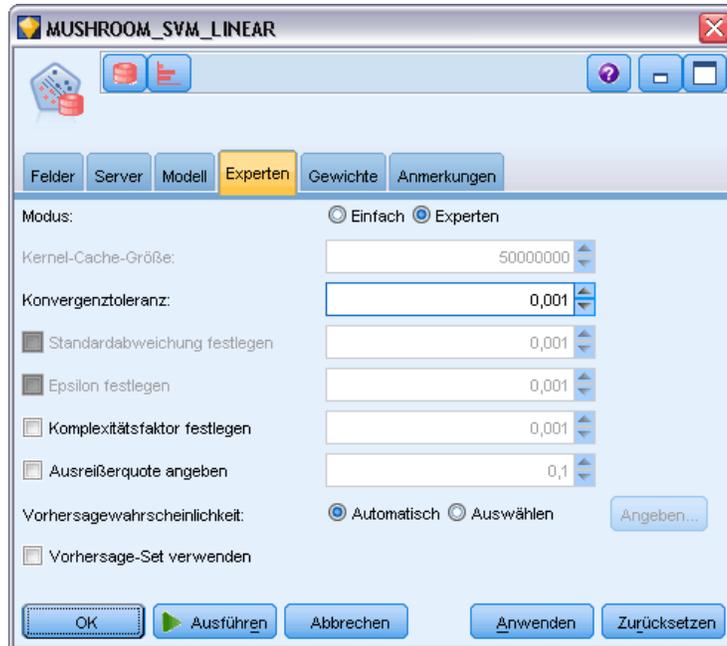
**Aktives Lernen.** Bietet eine Möglichkeit für den Umgang mit großen Aufbau-Sets. Beim aktiven Lernen erstellt der Algorithmus ein erstes Modell anhand einer kleinen Stichprobe, bevor er das Modell auf das gesamte Trainings-Daten-Set anwendet, und aktualisiert dann schrittweise die Stichprobe und das Modell anhand der Ergebnisse. Der Zyklus wird wiederholt, bis das Modell gegen die Trainingsdaten konvergiert oder bis die maximal zulässige Anzahl an Support-Vektoren erreicht wurde.

**Kernel-Funktion.** Wählen Sie Linear oder Gauss'sch oder belassen Sie den Standard Systembestimmt, damit das System den geeignetsten Kernel wählt. Gauss'sche Kernels sind in der Lage, komplexere Beziehungen zu lernen, benötigen aber in der Regel mehr Rechenzeit. Sie können mit dem linearen Kernel beginnen und den Gauss'schen Kernel nur dann ausprobieren, wenn der lineare Kernel keine gute Anpassung findet. Dies passiert häufiger mit einem Regressionsmodell, bei dem sich die Auswahl des Kernels stärker auswirkt. Denken Sie außerdem daran, dass mit dem Gauss'schen Kernel erzeugte SVM-Modelle nicht in SPSS Modeler durchsucht werden können. Mit dem linearen Kernel erstellte Modelle können genauso in SPSS Modeler durchsucht werden wie Standardregressionsmodelle.

**Normalisierungsmethode.** Legt die Normalisierungsmethode für kontinuierliche Eingabe- und Zielfelder fest. Zur Auswahl stehen Z-Score, Min-Max oder Keine. Oracle führt die Normalisierung automatisch durch, wenn das Kontrollkästchen Automatische Datenvorbereitung aktiviert ist. Deaktivieren Sie dieses Kontrollkästchen, um die Normalisierungsmethode manuell auszuwählen.

## Expertenoptionen für Oracle SVM

Abbildung 4-11  
SVM-Expertenoptionen



**Kernel-Cache-Größe.** Legt die Größe des während der Modellbildung zum Speichern berechneter Kernels verwendbaren Caches in Byte fest. Es liegt auf der Hand, dass ein größerer Cache in der Regel zu einer schnelleren Modellbildung führt. Der Standardwert ist 50MB.

**Konvergenztoleranz.** Legt den zulässigen Toleranzwert für den Abschluss der Modellbildung fest. Der Wert muss zwischen 0 und 1 liegen. Der Standardwert ist 0,001. Höhere Werte führen zu schneller erstellten, aber weniger genauen Modellen.

**Standardabweichung festlegen.** Legt den Standardabweichungsparameter fest, der vom Gauss'schen Kernel verwendet wird. Dieser Parameter wirkt sich auf das Verhältnis zwischen Modellkomplexität und der Möglichkeit der Generalisierung auf andere Daten-Sets aus (zu große und zu geringe Datenanpassung). Ein höherer Standardabweichungswert begünstigt eine zu geringe Anpassung. Standardmäßig wird dieser Parameter anhand der Trainingsdaten geschätzt.

**Epsilon festlegen.** Nur für Regressionsmodelle. Legt den Wert des Intervalls der Fehler fest, die bei der Bildung nicht Epsilon-sensitiver Modelle zulässig sind. Letztlich wird dadurch zwischen kleinen Fehlern (die ignoriert werden) und großen Fehlern (die nicht ignoriert werden) unterschieden. Der Wert muss zwischen 0 und 1 liegen. Standardmäßig wird dieser Wert aus den Trainingsdaten geschätzt.

**Komplexitätsfaktor festlegen.** Legt den Komplexitätsfaktor fest, der für den Ausgleich zwischen Modellfehler (gegen die Trainingsdaten gemessen) und Modellkomplexität sorgt und so eine zu große oder zu geringe Anpassung der Daten vermeidet. Ein höherer Wert stuft Fehler schwerwiegender ein, was das Risiko einer zu großen Anpassung der Daten birgt. Ein geringerer Wert stuft Fehler weniger schwerwiegend ein und kann zu einer geringen Anpassung führen.

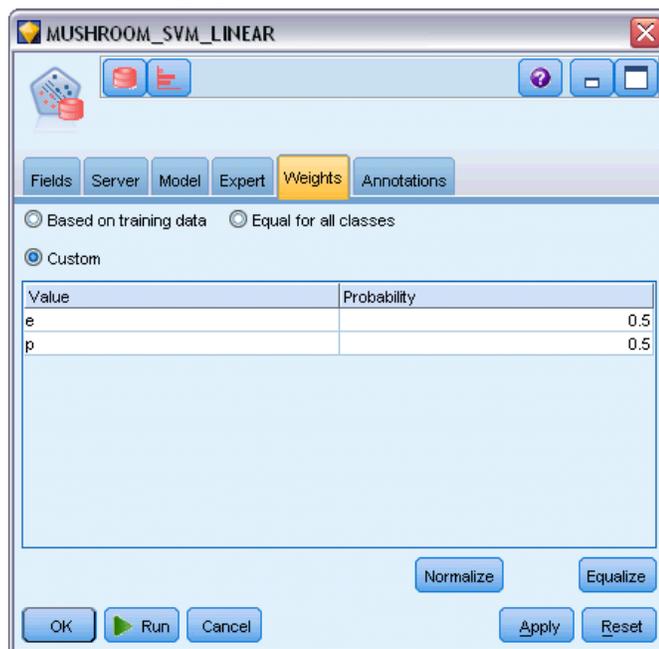
**Ausreißerquote angeben.** Gibt die gewünschte Quote an Ausreißern in den Trainingsdaten an. Nur gültig für SVM-Modellen mit einer einzigen Klasse. Die Verwendung mit der Einstellung **Komplexitätsfaktor festlegen** ist nicht möglich.

**Vorhersagewahrscheinlichkeit.** Ermöglicht dem Modell, die Wahrscheinlichkeit einer korrekten Prognose für ein mögliches Ergebnis des Zielfelds zu umfassen. Sie aktivieren diese Option, indem Sie Auswählen wählen, auf die Schaltfläche Angeben klicken, eines der möglichen Ergebnisse wählen und auf Einfügen klicken.

**Vorhersage-Set verwenden.** Generiert eine Tabelle von allen möglichen Resultaten für alle möglichen Ergebnisse des Zielfelds.

## Gewichtungsoptionen für Oracle SVM

Abbildung 4-12  
SVM-Gewichtungsoptionen



In einem Klassifikationsmodell können Sie mithilfe von Gewichten die relative Wichtigkeit der verschiedenen möglichen Zielwerte angeben. Dies kann beispielsweise dann sinnvoll sein, wenn die Datenpunkte in den Trainingsdaten nicht realistisch auf die verschiedenen Kategorien verteilt sind. Mit Gewichten können Sie das Modell verzerren, um einen Ausgleich für diejenigen Kategorien zu bewirken, die in den Daten unterrepräsentiert sind. Durch die Erhöhung des Gewichts für einen Zielwert sollte der Prozentsatz der richtigen Vorhersagen für die betreffende Kategorie erhöht werden.

Es gibt drei Methoden zur Festlegung von Gewichten:

- **Auf Trainingsdaten basierend.** Dies ist die Standardeinstellung. Die Gewichte basieren auf den relativen Häufigkeiten der Kategorien in den Trainingsdaten.

- **Für alle Klassen gleich.** Gewichte für alle Kategorien werden als  $1/k$  definiert, wobei  $k$  die Zahl der Zielkategorien darstellt.
- **Angepasst.** Sie können eigene Gewichte angeben. Die Startwerte für Gewichte werden für alle Klassen gleich gesetzt. Sie können die Gewichte für einzelne Kategorien auf benutzerdefinierte Werte einstellen. Um das Gewicht einer bestimmten Kategorie anzupassen, wählen Sie in der Tabelle die Gewichtungszelle aus, die der gewünschten Kategorie entspricht, löschen den Inhalt der Zelle und geben den gewünschten Wert ein.

Die Summe der Gewichte aller Kategorien sollte den Wert 1,0 ergeben. Wenn sie keine Summe von 1,0 bilden, wird eine Warnmeldung ausgegeben und es besteht die Möglichkeit, die Werte automatisch normalisieren zu lassen. Diese automatische Anpassung behält die Anteile über die Kategorien hinweg bei, während die Gewichtsbeschränkung erzwungen wird. Sie können diese Anpassung jederzeit durchführen, indem Sie auf die Schaltfläche **Normalisieren** klicken. Um die Tabelle auf gleiche Werte für alle Kategorien zurückzusetzen, klicken Sie auf die Schaltfläche **Gleichsetzen**.

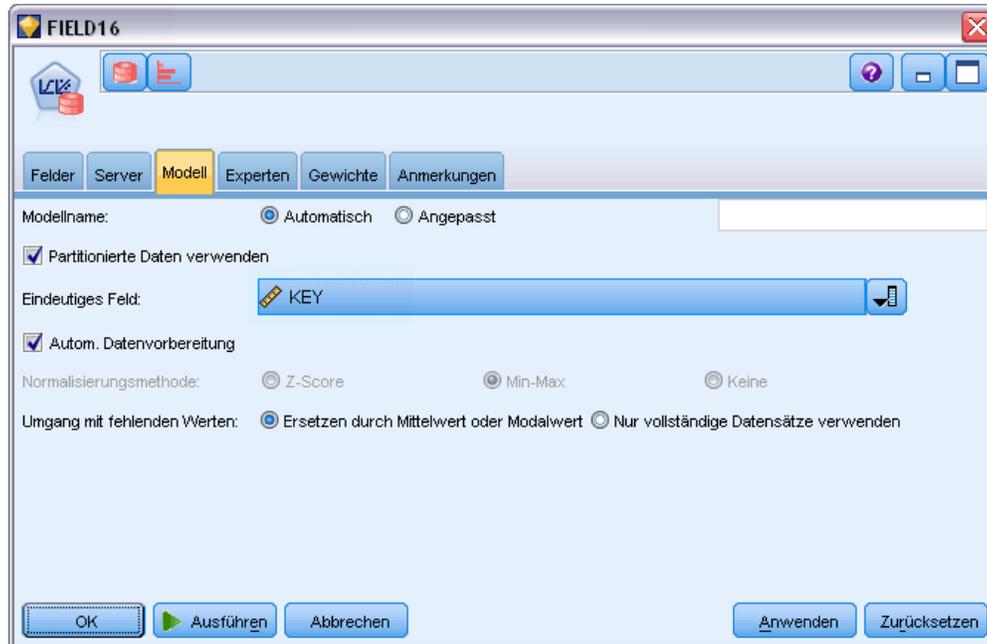
## **Verallgemeinerte lineare Modelle (GLM) von Oracle**

(Nur 11g) Verallgemeinerte lineare Modelle lockern die restriktiven Annahmen der linearen Modelle. Dazu gehören beispielsweise die Annahmen, dass die Zielvariable eine Normalverteilung hat und dass die Wirkung der Prädiktoren auf die Zielvariable in ihrem Wesen linear ist. Ein verallgemeinertes lineares Modell eignet sich für Vorhersagen, in denen das Ziel wahrscheinlich eine nichtnormale Verteilung hat, z. B. eine Multinomial- oder eine Poisson-Verteilung. Ebenso ist ein verallgemeinertes lineares Modell nützlich, wenn die Beziehung oder die Verknüpfung zwischen den Prädiktoren und dem Ziel wahrscheinlich nicht linear ist.

Weitere Informationen finden Sie in folgenden Publikationen: *Oracle Data Mining Application Developer's Guide* und *Oracle Data Mining Concepts*.

## Optionen für Oracle GLM-Modelle

Abbildung 4-13  
GLM-Modelloptionen



**Modellname.** Sie können den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen (oder dem Modelltyp in Fällen, in denen kein solches Feld angegeben ist) generieren oder einen benutzerdefinierten Namen eingeben.

**Eindeutiges Feld.** Gibt das Feld an, über das jeder Fall eindeutig identifiziert wird. Dies kann beispielsweise ein ID-Feld sein, wie *CustomerID*. IBM® SPSS® Modeler schreibt vor, dass dieses Schlüsselfeld zwingend numerisch sein muss.

*Anmerkung:* Dieses Feld ist für alle Oracle-Knoten mit Ausnahme von Oracle Adaptive Bayes, Oracle O-Cluster und Oracle Apriori optional.

**Automatische Datenvorbereitung.** (Nur 11g only) Aktiviert (Standard) oder deaktiviert den automatisierten Datenvorbereitungsmodus von Oracle Data Mining. Wenn dieses Kästchen markiert ist, führt ODM automatisch die vom Algorithmus geforderten Datenumformungen durch. Weitere Informationen finden Sie unter *Oracle Data Mining Concepts*.

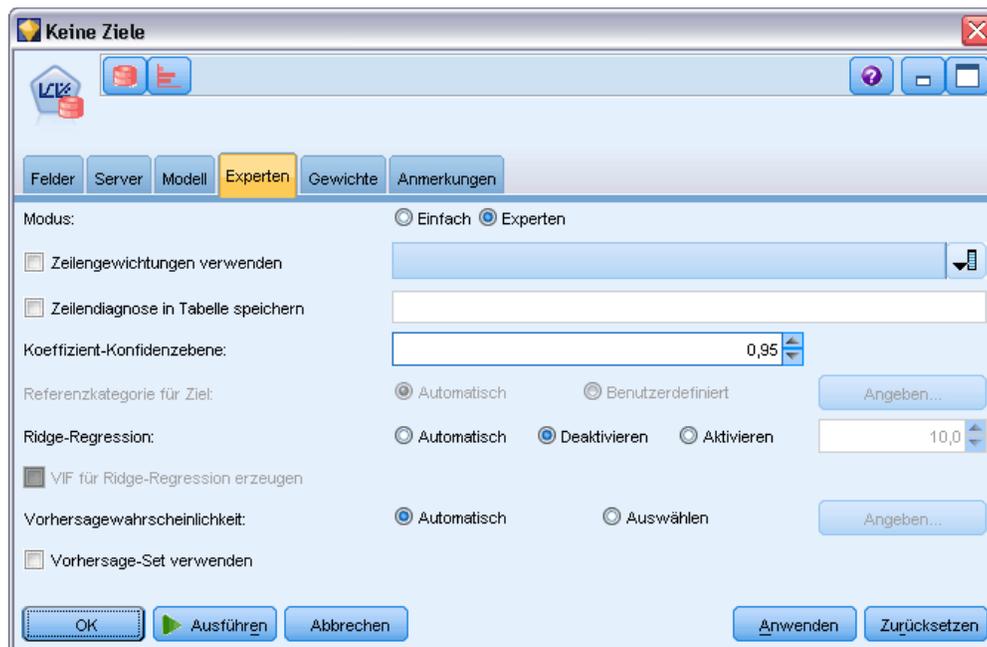
**Normalisierungsmethode.** Legt die Normalisierungsmethode für kontinuierliche Eingabe- und Zielfelder fest. Zur Auswahl stehen Z-Score, Min-Max oder Keine. Oracle führt die Normalisierung automatisch durch, wenn das Kontrollkästchen Automatische Datenvorbereitung aktiviert ist. Deaktivieren Sie dieses Kontrollkästchen, um die Normalisierungsmethode manuell auszuwählen.

**Umgang mit fehlenden Werten.** Gibt an, wie fehlende Werte in den Eingabedaten verarbeitet werden sollen:

- Ersetzen durch Mittelwert oder Modalwert ersetzt fehlende Werte von numerischen Attributen durch den Mittelwert und fehlende Werte von kategorialen Attributen durch den Modalwert.
- Nur vollständige Datensätze verwenden ignoriert Datensätze, in denen Werte fehlen.

## Expertenoptionen für Oracle GLM

Abbildung 4-14  
GLM-Expertenoptionen



**Zeilengewichtungen verwenden.** Markieren Sie dieses Kontrollkästchen, um die benachbarte Dropdown-Liste zu aktivieren, aus der Sie eine Spalte mit einem Gewichtungsfaktor für die Zeilen wählen können.

**Zeilendiagnose in Tabelle speichern.** Markieren Sie dieses Kontrollkästchen, um das benachbarte Textfeld zu aktivieren, in das Sie den Namen einer Tabelle eingeben können, die Diagnosedaten auf Zeilenebene enthalten soll.

**Koeffizient-Konfidenzebene.** Der Sicherheitsgrad (von 0.0 bis 1.0), in dem der vorhergesagte Wert für das Ziel innerhalb eines Konfidenzintervalls liegt, das vom Modell berechnet wurde. Konfidenzgrenzen werden mit den Koeffizientenstatistiken zurückgegeben.

**Referenzkategorie für Ziel.** Wählen Sie Benutzerdefiniert aus, um für das Zielfeld einen Wert als Referenzkategorie zu verwenden, oder behalten Sie den Standardwert Auto.

**Ridge-Regression.** Bei der Ridge-Regression handelt es sich um eine Technik zur Kompensierung der Situation, in der ein zu hoher Korrelationsgrad bei den Variablen besteht. Mithilfe der Option Auto können Sie erlauben, dass der Algorithmus diese Technik verwendet. Sie können sie aber auch manuell über die Optionen Deaktivieren und Aktivieren steuern. Wenn Sie sich für die manuelle

Aktivierung der Ridge-Regression entscheiden, können Sie den Standardwert des Systems für den Ridge-Parameter überschreiben, indem Sie einen Wert in das benachbarte Feld eingeben.

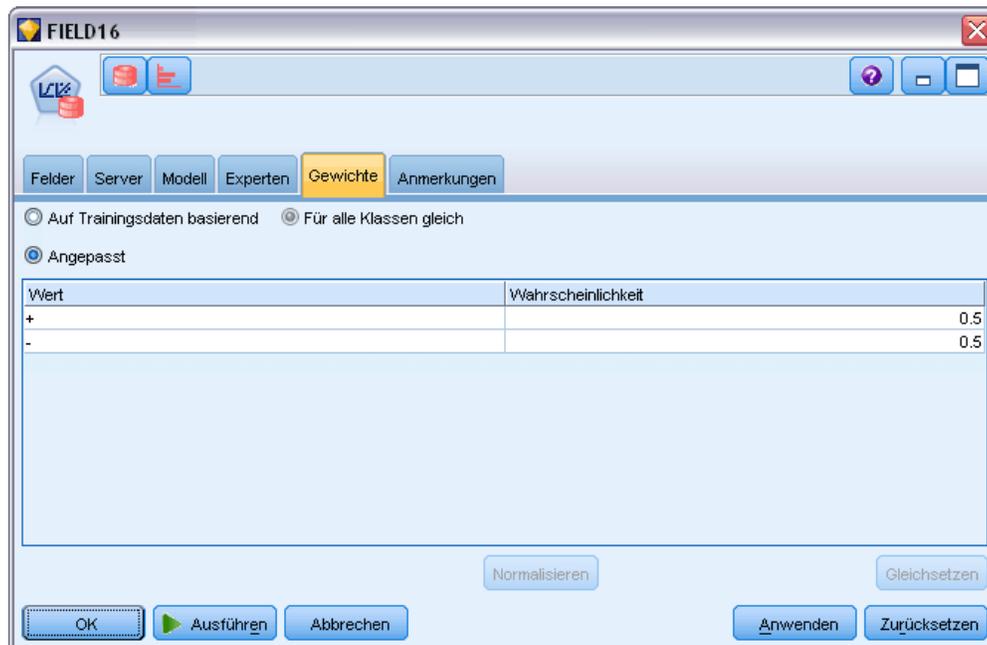
**VIF für Ridge-Regression erzeugen.** Markieren Sie dieses Kontrollkästchen zur Erzeugung von VIF-Statistiken (Variance Inflation Factor), wenn die Ridge für lineare Regression benutzt wird.

**Vorhersagewahrscheinlichkeit.** Ermöglicht dem Modell, die Wahrscheinlichkeit einer korrekten Prognose für ein mögliches Ergebnis des Zielfelds zu umfassen. Sie aktivieren diese Option, indem Sie Auswählen wählen, auf die Schaltfläche Angeben klicken, eines der möglichen Ergebnisse wählen und auf Einfügen klicken.

**Vorhersage-Set verwenden.** Generiert eine Tabelle von allen möglichen Resultaten für alle möglichen Ergebnisse des Zielfelds.

## Gewichtungsoptionen für Oracle GLM

Abbildung 4-15  
GLM-Gewichtungsoptionen



In einem Klassifikationsmodell können Sie mithilfe von Gewichten die relative Wichtigkeit der verschiedenen möglichen Zielwerte angeben. Dies kann beispielsweise dann sinnvoll sein, wenn die Datenpunkte in den Trainingsdaten nicht realistisch auf die verschiedenen Kategorien verteilt sind. Mit Gewichten können Sie das Modell verzerren, um einen Ausgleich für diejenigen Kategorien zu bewirken, die in den Daten unterrepräsentiert sind. Durch die Erhöhung des Gewichts für einen Zielwert sollte der Prozentsatz der richtigen Vorhersagen für die betreffende Kategorie erhöht werden.

Es gibt drei Methoden zur Festlegung von Gewichten:

- **Auf Trainingsdaten basierend.** Dies ist die Standardeinstellung. Die Gewichte basieren auf den relativen Häufigkeiten der Kategorien in den Trainingsdaten.

- **Für alle Klassen gleich.** Gewichte für alle Kategorien werden als  $1/k$  definiert, wobei  $k$  die Zahl der Zielkategorien darstellt.
- **Angepasst.** Sie können eigene Gewichte angeben. Die Startwerte für Gewichte werden für alle Klassen gleich gesetzt. Sie können die Gewichte für einzelne Kategorien auf benutzerdefinierte Werte einstellen. Um das Gewicht einer bestimmten Kategorie anzupassen, wählen Sie in der Tabelle die Gewichtungszelle aus, die der gewünschten Kategorie entspricht, löschen den Inhalt der Zelle und geben den gewünschten Wert ein.

Die Summe der Gewichte aller Kategorien sollte den Wert 1,0 ergeben. Wenn sie keine Summe von 1,0 bilden, wird eine Warnmeldung ausgegeben und es besteht die Möglichkeit, die Werte automatisch normalisieren zu lassen. Diese automatische Anpassung behält die Anteile über die Kategorien hinweg bei, während die Gewichtsbeschränkung erzwungen wird. Sie können diese Anpassung jederzeit durchführen, indem Sie auf die Schaltfläche **Normalisieren** klicken. Um die Tabelle auf gleiche Werte für alle Kategorien zurückzusetzen, klicken Sie auf die Schaltfläche **Gleichsetzen**.

## **Oracle Decision Tree**

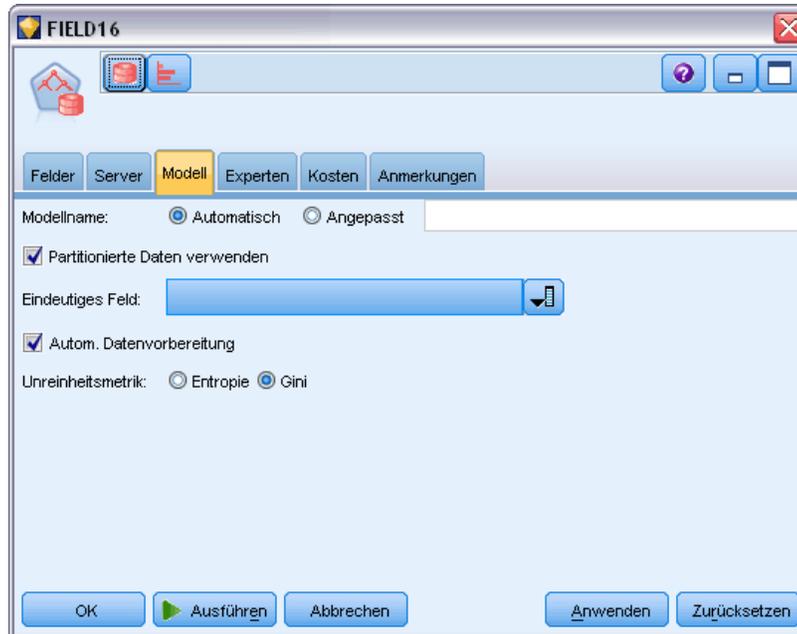
Oracle Data Mining bietet eine klassische Entscheidungsbaumfunktion, die auf dem beliebten Algorithmus mit Klassifizierungs- und Regressions-Bäumen beruht. Das ODM Decision Tree-Modell enthält vollständige Informationen zu jedem Knoten, einschließlich Konfidenz, Support und Splitting-Kriterium. Für jeden Knoten kann die vollständige Regel angezeigt werden. Außerdem wird für jeden Knoten ein Ersatzattribut angegeben, das verwendet wird, wenn das Modell auf einen Fall mit fehlenden Werten angewendet wird.

Entscheidungsbäume sind beliebt, weil sie universell einsetzbar, leicht anzuwenden und leicht zu verstehen sind. Entscheidungsbäume sichten alle potenziellen Eingabeattribute auf der Suche nach dem besten "Splitter", also dem besten Trennwert für Attribute (z. B.  $AGE > 55$ ), der die weiter unten im Stream liegenden Datensätze in homogenere Grundgesamtheiten aufteilt. Bei jeder Split-Entscheidung wiederholt ODM den Vorgang, indem der gesamte Baum erweitert wird und End-"Blätter" erstellt werden, die ähnliche Grundgesamtheiten von Datensätzen, Elementen bzw. Personen darstellen. Ausgehend vom Stammknoten des Baums (z. B. der gesamten Grundgesamtheit) bieten Entscheidungsbäume von Menschen lesbare Regeln mit Anweisungen vom Typ **IF A, then B**. Diese Entscheidungsbaumregeln geben außerdem Support und Konfidenz für jeden Baumknoten an.

Auch Adaptive Bayes Networks können kurze und einfache Regeln bieten, die dazu beitragen können, Erklärungen für jede Vorhersage zu finden, Entscheidungsbäume jedoch bieten vollständige Oracle Data Mining-Regeln für jede Split-Entscheidung. Entscheidungsbäume sind außerdem hilfreich bei der Entwicklung detaillierter Profile für die besten Kunden, gesunde Patienten, Faktoren im Zusammenhang mit Betrug usw.

## Optionen für Entscheidungsbaummodelle

Abbildung 4-16  
Optionen für Entscheidungsbaummodelle



**Modellname.** Sie können den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen (oder dem Modelltyp in Fällen, in denen kein solches Feld angegeben ist) generieren oder einen benutzerdefinierten Namen eingeben.

**Eindeutiges Feld.** Gibt das Feld an, über das jeder Fall eindeutig identifiziert wird. Dies kann beispielsweise ein ID-Feld sein, wie *CustomerID*. IBM® SPSS® Modeler schreibt vor, dass dieses Schlüsselfeld zwingend numerisch sein muss.

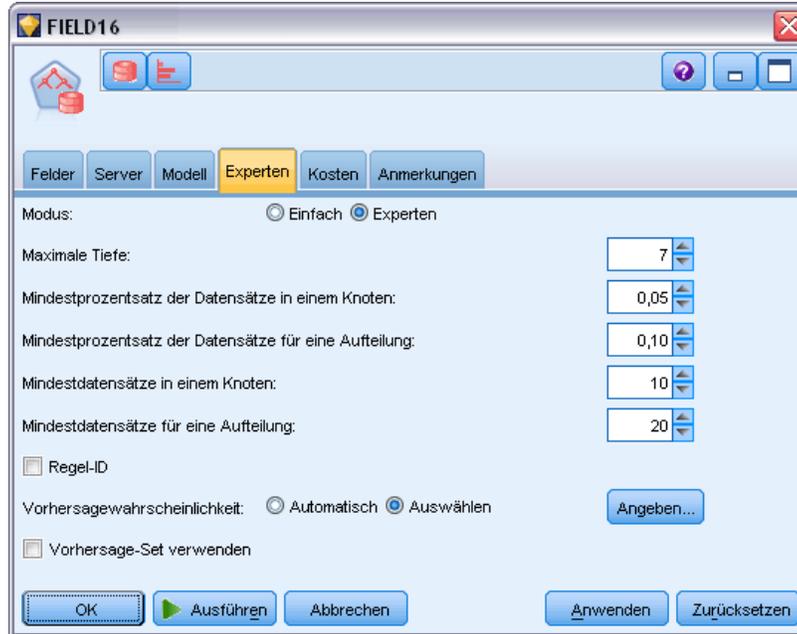
*Anmerkung:* Dieses Feld ist für alle Oracle-Knoten mit Ausnahme von Oracle Adaptive Bayes, Oracle O-Cluster und Oracle Apriori optional.

**Automatische Datenvorbereitung.** (Nur 11g only) Aktiviert (Standard) oder deaktiviert den automatisierten Datenvorbereitungsmodus von Oracle Data Mining. Wenn dieses Kästchen markiert ist, führt ODM automatisch die vom Algorithmus geforderten Datenumformungen durch. Weitere Informationen finden Sie unter *Oracle Data Mining Concepts*.

**Unreinheitsmetrik.** Gibt an, welche Metrik für die Ermittlung der besten Testfrage für die Aufteilung der Daten an den einzelnen Knoten verwendet wird. Der beste Splitter und Split-Wert sind diejenigen, die zu der größten Zunahme an Zielwerthomogenität für die Elemente im Knoten führen. Homogenität wird in Übereinstimmung mit einer Metrik gemessen. Die Metriken **Gini** und **Entropie** werden unterstützt.

## Expertenoptionen für Entscheidungsbäume

Abbildung 4-17  
Expertenoptionen für Entscheidungsbäume



**Maximale Tiefe.** Legt die maximale Tiefe des zu erstellenden Baummodells fest.

**Mindestprozentatz der Datensätze in einem Knoten.** Legt den Prozentsatz der Mindestanzahl an Datensätzen pro Knoten fest.

**Mindestprozentatz der Datensätze für eine Aufteilung.** Legt die Mindestanzahl an Datensätzen in einem übergeordneten Knoten als Prozentsatz der Gesamtzahl der zum Trainieren des Modells verwendeten Datensätze fest. Es wird nicht versucht, einen Split durchzuführen, wenn die Anzahl der Datensätze unterhalb dieses Prozentsatzes liegt.

**Mindestdatensätze in einem Knoten.** Legt die Mindestanzahl an auszugebenden Datensätzen fest.

**Mindestdatensätze für eine Aufteilung.** Legt die Mindestanzahl der Datensätze in einem übergeordneten Knoten als Wert fest. Es wird nicht versucht, einen Split durchzuführen, wenn die Anzahl der Datensätze unterhalb dieses Werts liegt.

**Regel-ID.** Wenn diese Option aktiviert ist, wird eine Zeichenkette in das Modell aufgenommen, die den Knoten im Baum angibt, bei dem eine bestimmte Aufteilung (Split) vorgenommen werden soll.

**Vorhersagewahrscheinlichkeit.** Ermöglicht dem Modell, die Wahrscheinlichkeit einer korrekten Prognose für ein mögliches Ergebnis des Zielfelds zu umfassen. Sie aktivieren diese Option, indem Sie Auswählen wählen, auf die Schaltfläche Angeben klicken, eines der möglichen Ergebnisse wählen und auf Einfügen klicken.

**Vorhersage-Set verwenden.** Generiert eine Tabelle von allen möglichen Resultaten für alle möglichen Ergebnisse des Zielfelds.

## Oracle O-Cluster

Der Algorithmus “Oracle O-Cluster” identifiziert natürlich vorkommende Gruppierungen in einer Datengesamtheit. Clustering mit orthogonaler Partitionierung (O-Cluster) ist ein Oracle-eigener Clustering-Algorithmus, der ein auf einem hierarchischen Gitter beruhendes Clustering-Modell erstellt, d. h., es erstellt achsenparallele (orthogonale) Partitionen im Bereich des Eingabeattributraums. Der Algorithmus arbeitet rekursiv. Die entstehende hierarchische Struktur stellt ein unregelmäßiges Gitter dar, das den Attributraum in Cluster zerlegt.

Der O-Cluster-Algorithmus kann sowohl numerische als auch kategoriale Attribute verarbeiten und ODM wählt automatisch die besten Cluster-Definitionen aus. ODM gibt Informationen zu Cluster-Details, Cluster-Regeln, Werte für den Cluster-Schwerpunkt (Zentroid) und kann zum Scoren einer Grundgesamtheit in Bezug auf ihre Cluster-Zugehörigkeit verwendet werden.

### Modelloptionen für O-Cluster

Abbildung 4-18  
Modelloptionen für O-Cluster



**Modellname.** Sie können den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen (oder dem Modelltyp in Fällen, in denen kein solches Feld angegeben ist) generieren oder einen benutzerdefinierten Namen eingeben.

**Eindeutiges Feld.** Gibt das Feld an, über das jeder Fall eindeutig identifiziert wird. Dies kann beispielsweise ein ID-Feld sein, wie *CustomerID*. IBM® SPSS® Modeler schreibt vor, dass dieses Schlüsselfeld zwingend numerisch sein muss.

*Anmerkung:* Dieses Feld ist für alle Oracle-Knoten mit Ausnahme von Oracle Adaptive Bayes, Oracle O-Cluster und Oracle Apriori optional.

**Automatische Datenvorbereitung.** (Nur 11g only) Aktiviert (Standard) oder deaktiviert den automatisierten Datenvorbereitungsmodus von Oracle Data Mining. Wenn dieses Kästchen markiert ist, führt ODM automatisch die vom Algorithmus geforderten Datenumformungen durch. Weitere Informationen finden Sie unter *Oracle Data Mining Concepts*.

**Maximale Anzahl von Clustern.** Legt die maximale Anzahl der generierten Cluster fest.

## Expertenoptionen für O-Cluster

Abbildung 4-19  
Expertenoptionen für O-Cluster



**Maximaler Puffer.** Legt die maximale Puffergröße fest.

**Sensitivität.** Legt einen Anteil fest, der die für die Abtrennung eines neuen Clusters erforderliche Spitzendichte angibt. Der Anteil steht in Bezug zur globalen einheitlichen Dichte.

## Oracle k-Means

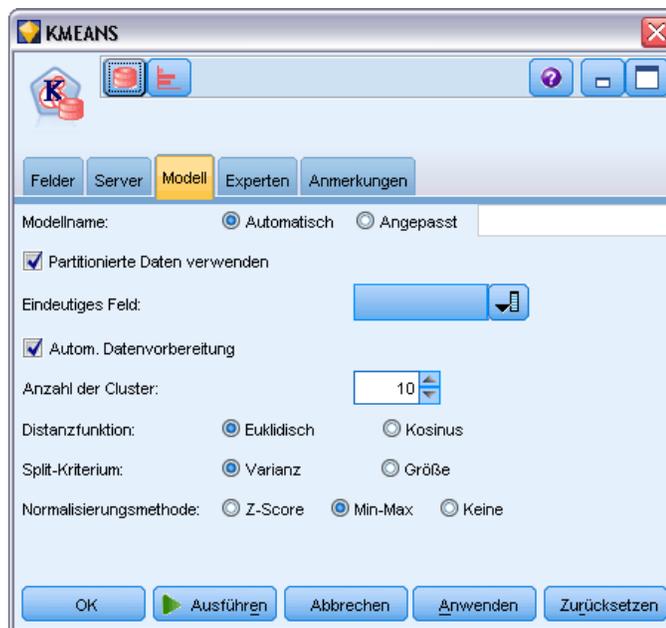
Der Algorithmus "Oracle k-Means" identifiziert natürlich vorkommende Gruppierungen in einer Datengesamtheit. Der k-Means-Algorithmus ist ein distanzbasierter Cluster-Algorithmus, der die Daten in eine zuvor festgelegte Anzahl an Clustern einteilt (vorausgesetzt, dass genügend unterschiedliche Fälle vorhanden sind). Distanzbasierte Algorithmen beruhen auf einer Distanzmetrik (Funktion) zur Messung der Ähnlichkeit zwischen Datenpunkten. Datenpunkte werden dem nächsten Cluster gemäß der verwendeten Distanzmetrik zugewiesen. ODM bietet eine erweiterte Version von k-Means.

Der k-Means-Algorithmus unterstützt hierarchische Cluster, verarbeitet numerische und kategoriale Attribute und teilt die Grundgesamtheit in die vom Benutzer angegebene Anzahl an Clustern auf. ODM gibt Informationen zu Cluster-Details, Cluster-Regeln, Werte für den

Cluster-Schwerpunkt (Zentroid) und kann zum Scoring einer Grundgesamtheit in Bezug auf ihre Cluster-Zugehörigkeit verwendet werden.

## Optionen für das k-Means-Modell

Abbildung 4-20  
Optionen für das k-Means-Modell



**Modellname.** Sie können den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen (oder dem Modelltyp in Fällen, in denen kein solches Feld angegeben ist) generieren oder einen benutzerdefinierten Namen eingeben.

**Eindeutiges Feld.** Gibt das Feld an, über das jeder Fall eindeutig identifiziert wird. Dies kann beispielsweise ein ID-Feld sein, wie *CustomerID*. IBM® SPSS® Modeler schreibt vor, dass dieses Schlüsselfeld zwingend numerisch sein muss.

*Anmerkung:* Dieses Feld ist für alle Oracle-Knoten mit Ausnahme von Oracle Adaptive Bayes, Oracle O-Cluster und Oracle Apriori optional.

**Automatische Datenvorbereitung.** (Nur 11g only) Aktiviert (Standard) oder deaktiviert den automatisierten Datenvorbereitungsmodus von Oracle Data Mining. Wenn dieses Kästchen markiert ist, führt ODM automatisch die vom Algorithmus geforderten Datenumformungen durch. Weitere Informationen finden Sie unter *Oracle Data Mining Concepts*.

**Anzahl der Cluster.** Legt die Anzahl der generierten Cluster fest.

**Distanzfunktion.** Gibt an, welche Distanzfunktion für k-Means-Clustering verwendet wird.

**Split-Kriterium.** Gibt an, welches Split-Kriterium für k-Means-Clustering verwendet wird.

**Normalisierungsmethode.** Legt die Normalisierungsmethode für kontinuierliche Eingabe- und Zielfelder fest. Zur Auswahl stehen Z-Score, Min-Max oder Keine.

## K-Means-Expertenoptionen

Abbildung 4-21  
K-Means-Expertenoptionen



**Iterationen.** Legt die Anzahl der Iterationen für den k-Means-Algorithmus fest.

**Konvergenztoleranz.** Legt die Konvergenztoleranz für den k-Means-Algorithmus fest.

**Anzahl der Klassen.** Gibt die Anzahl der Klassen in dem von k-Means erstellten Attributhistogramm an. Die Klassengrenzen für die einzelnen Attribute werden global für das gesamte Trainingsdaten-Set berechnet. Die Klassiermethode ist "Gleiche Breite". Alle Attribute haben dieselbe Anzahl von Klassen, mit Ausnahme von Attributen mit einem einzelnen Wert, die nur eine einzige Klasse aufweisen.

**Blockerweiterung.** Legt den Erweiterungsfaktor für Arbeitsspeicher fest, der für die Aufnahme der Clusterdaten zugewiesen wird.

**Mindestprozentsatz für Attribut-Support.** Legt den Anteil der Attributwerte fest, die nicht null sein müssen, damit das Attribut in die Regelbeschreibung für den Cluster aufgenommen wird. Wenn der Parameterwert bei Daten mit fehlenden Werten zu hoch festgelegt wird, kann dies zu sehr kurzen oder sogar leeren Regeln führen.

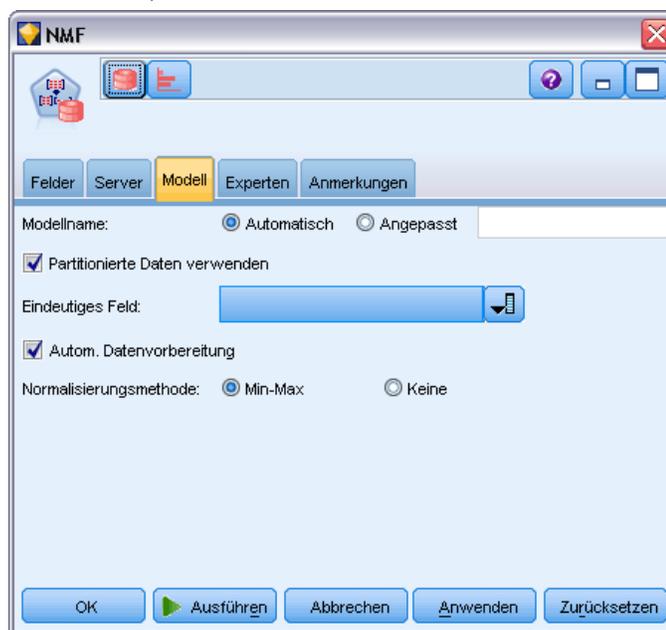
## Oracle Nonnegative Matrix Factorization (NMF)

"Nonnegative Matrix Factorization" (NMF) dient zur Verkleinerung eines großen Daten-Sets in repräsentative Attribute. NMF ähnelt vom Konzept her der Hauptkomponentenanalyse (Principal Components Analysis, PCA), kann jedoch mit größeren Attributmengen und einem additiven Darstellungsmodell umgehen und ist somit ein leistungsstarker, hochmoderner Data-Mining-Algorithmus, der für eine Vielzahl von Verwendungsfällen eingesetzt werden kann.

NMF kann verwendet werden, um große Datenmengen, beispielsweise Textdaten, in kleinere, dünner besetzte Darstellungen zu reduzieren, die die Dimensionalität der Daten verringern (dieselben Informationen können unter Verwendung von wesentlich weniger Variablen beibehalten werden). Die Ausgabe der NMF-Modelle kann mithilfe von Techniken für überwachtetes Lernen, wie SVMs, oder nicht überwachtetes Lernen, wie Clustering-Verfahren, analysiert werden. Oracle Data Mining verwendet NMF- und SVM-Algorithmen für das Mining von unstrukturierten Textdaten.

## NMF-Modelloptionen

Abbildung 4-22  
NMF-Modelloptionen



**Modellname.** Sie können den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen (oder dem Modelltyp in Fällen, in denen kein solches Feld angegeben ist) generieren oder einen benutzerdefinierten Namen eingeben.

**Eindeutiges Feld.** Gibt das Feld an, über das jeder Fall eindeutig identifiziert wird. Dies kann beispielsweise ein ID-Feld sein, wie *CustomerID*. IBM® SPSS® Modeler schreibt vor, dass dieses Schlüsselfeld zwingend numerisch sein muss.

*Anmerkung:* Dieses Feld ist für alle Oracle-Knoten mit Ausnahme von Oracle Adaptive Bayes, Oracle O-Cluster und Oracle Apriori optional.

**Automatische Datenvorbereitung.** (Nur 11g only) Aktiviert (Standard) oder deaktiviert den automatisierten Datenvorbereitungsmodus von Oracle Data Mining. Wenn dieses Kästchen markiert ist, führt ODM automatisch die vom Algorithmus geforderten Datenumformungen durch. Weitere Informationen finden Sie unter *Oracle Data Mining Concepts*.

**Normalisierungsmethode.** Legt die Normalisierungsmethode für kontinuierliche Eingabe- und Zielfelder fest. Zur Auswahl stehen Z-Score, Min-Max oder Keine. Oracle führt die Normalisierung automatisch durch, wenn das Kontrollkästchen Automatische Datenvorbereitung aktiviert ist. Deaktivieren Sie dieses Kontrollkästchen, um die Normalisierungsmethode manuell auszuwählen.

## NMF-Expertenoptionen

Abbildung 4-23  
NMF-Expertenoptionen



**Anzahl der Merkmale angeben.** Dient zur Angabe der Anzahl der zu extrahierenden Merkmale.

**Startwert für Zufallsgenerator.** Legt den Zufallsstartwert für den NMF-Algorithmus fest.

**Anzahl der Iterationen.** Legt die Anzahl der Iterationen für den NMF-Algorithmus fest.

**Konvergenztoleranz.** Legt die Konvergenztoleranz für den NMF-Algorithmus fest.

**Alle Merkmale anzeigen.** Zeigt die Merkmals-ID und die Konfidenz für alle Merkmale an und nicht nur die Werte für das beste Merkmal.

## Oracle Apriori

Der Algorithmus “A priori” erkennt Assoziationsregeln in den Daten. Beispiel: “Falls ein Kunde einen Rasierer und After-Shave-Lotion kauft, dann kauft er auch mit 80%iger Wahrscheinlichkeit Rasiercreme.” Das Association-Mining-Problem lässt sich in zwei Teilprobleme zerlegen:

- Ermitteln aller Elementkombinationen, der so genannten “Frequent Itemsets” (häufig vorkommende Elementmengen), deren Support größer ist als der minimale Support.
- Verwenden der Frequent Itemsets zum Generieren der gewünschten Regeln. Es gilt also: Wenn ABC und BC häufig vorkommen, dann gilt die Regel “A impliziert BC” immer dann, wenn das Verhältnis von  $\text{support}(ABC)$  zu  $\text{support}(BC)$  mindestens so groß ist wie die minimale Konfidenz. Beachten Sie, dass die Regel über den Mindestsupport verfügt, da ABCD häufig vorkommt. ODM Association unterstützt nur Regeln mit einem einzigen Sukzedens (ABC impliziert D).

Die Anzahl der Frequent Itemsets richtet sich nach den Parametern für den minimalen Support. Die Anzahl der generierten Regeln richtet sich nach der Anzahl der Frequent Itemsets und dem Konfidenzparameter. Wenn der Konfidenzparameter zu hoch festgelegt ist, kann es vorkommen, dass zwar Frequent Itemsets im Assoziationsmodell vorliegen, aber keine Regeln.

ODM verwendet eine SQL-basierte Implementierung des Algorithmus “A priori” Die Schritte zur Kandidatengenerierung und zur Support-Zählung werden mithilfe von SQL-Abfragen implementiert. Es werden keine spezialisierten, arbeitsspeicherinternen Datenstrukturen verwendet. Die SQL-Abfragen werden durch verschiedene Hinweise so optimiert, dass sie effizient auf dem Datenbankserver ausgeführt werden.

## Feldoptionen für A Priori

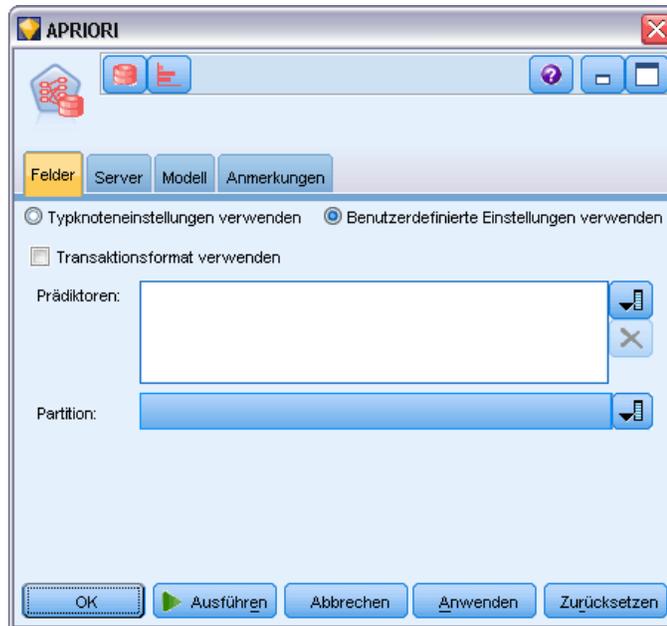
Alle Modellierungsknoten besitzen die Registerkarte “Felder”, auf der Sie die Felder festlegen können, die beim Erstellen des Modells verwendet werden.

Bevor Sie ein “A Priori”-Model erstellen können, müssen Sie festlegen, welche Felder als relevante Elemente bei der Assoziationsmodellierung verwendet werden sollen.

**Typknoteneinstellungen verwenden.** Diese Option weist den Knoten an, die Feldinformationen von einem weiter oben liegenden Typknoten zu verwenden. Dies ist die Standardeinstellung.

**Benutzerdefinierte Einstellungen verwenden.** Diese Option weist den Knoten an, die hier angegebenen Feldinformationen anstelle der in einem weiter oben liegenden Typknoten angegebenen zu verwenden. Geben Sie nach Auswahl dieser Option die restlichen Felder im Dialogfeld an. (Diese hängen davon ab, ob Sie das Transaktionsformat verwenden.)

Abbildung 4-24  
Standardeinstellungen für benutzerdefinierte Felder



Wenn Sie das Transaktionsformat *nicht verwenden*, geben Sie Folgendes an:

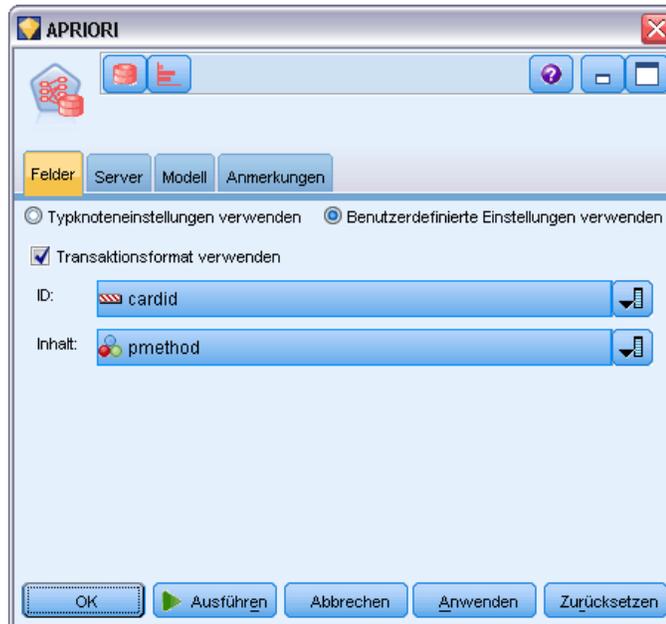
- **Eingaben.** Wählen Sie die Eingabefelder aus. Dies ist so, als würden Sie in einem Typknoten für die Rolle eines Felds den Wert *Eingabe* festlegen.
- **Partition.** In diesem Feld können Sie ein Feld angeben, das verwendet wird, um die Daten in getrennte Stichproben für die Trainings-, Test- und Validierungsphase der Modellbildung aufzuteilen. [Für weitere Informationen siehe Thema Partitionsknoten in Kapitel 4 in IBM SPSS Modeler 14.2- Quellen-, Prozess- und Ausgabeknoten.](#)

Wenn Sie das Transaktionsformat *verwenden*, geben Sie Folgendes an:

**Transaktionsformat verwenden.** Verwenden Sie diese Option, wenn Sie Daten so transformieren möchten, dass nicht mehr eine Zeile pro Element, sondern eine Zeile pro Fall verwendet wird.

Durch Auswahl dieser Option werden die Feld-Steuerelemente im unteren Bereich dieses Dialogfelds verändert:

Abbildung 4-25  
Feldeinstellungen für Transaktionsformat

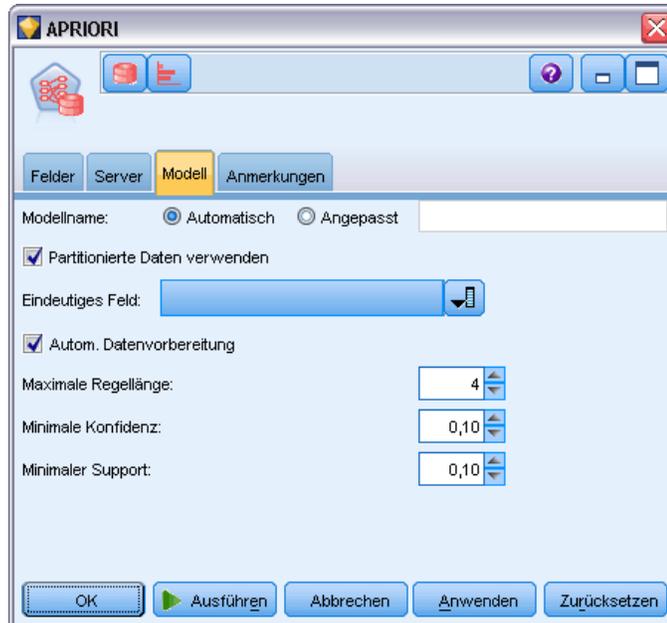


Geben Sie beim Transaktionsformat Folgendes an:

- **ID.** Wählen Sie ein ID-Feld aus der Liste aus. Als ID-Feld können numerische oder symbolische Felder verwendet werden. Jeder eindeutige Wert in diesem Feld sollte eine bestimmte Analyseeinheit darstellen. Bei einer Warenkorbanwendung könnte z. B. jede ID einen einzelnen Kunden darstellen. Für eine Webprotokoll-Analyseanwendung könnte jede ID einen Computer (nach IP-Adresse) oder einen Benutzer (nach Anmeldedaten) darstellen.
- **Inhalt.** Geben Sie das Inhaltsfeld für das Modell an. Dieses Feld enthält das Element, das für die Assoziationsmodellierung relevant ist.
- **Partition.** In diesem Feld können Sie ein Feld angeben, das verwendet wird, um die Daten in getrennte Stichproben für die Trainings-, Test- und Validierungsphase der Modellbildung aufzuteilen. Indem Sie mit einer Stichprobe das Modell erstellen und es mit einer anderen Stichprobe testen, erhalten Sie einen guten Hinweis dafür, wie gut das Modell sich für größere Datenmengen generalisieren lässt, die den aktuellen Daten ähneln. Wenn mehrere Partitionsfelder mithilfe von Typ- oder Partitionsknoten erstellt wurden, muss in jedem Modellierungsknoten, der die Partitionierung verwendet, auf der Registerkarte "Felder" ein einzelnes Partitionsfeld ausgewählt werden. (Wenn nur eine einzige Partition vorhanden ist, wird diese immer automatisch verwendet, wenn die Partitionierung aktiviert ist.) [Für weitere Informationen siehe Thema Partitionsknoten in Kapitel 4 in IBM SPSS Modeler 14.2-Quellen-, Prozess- und Ausgabeknoten.](#) Beachten Sie außerdem, dass die Partitionierung auch auf der Registerkarte "Modelloptionen" des Knotens aktiviert sein muss, wenn die ausgewählte Partitionierung in Ihrer Analyse angewendet werden soll. (Wenn die Auswahl der Option aufgehoben ist, kann die Partitionierung ohne Änderung der Feldeinstellungen deaktiviert werden.)

## Modelloptionen für A Priori

Abbildung 4-26  
Modelloptionen für A Priori



**Modellname.** Sie können den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen (oder dem Modelltyp in Fällen, in denen kein solches Feld angegeben ist) generieren oder einen benutzerdefinierten Namen eingeben.

**Eindeutiges Feld.** Gibt das Feld an, über das jeder Fall eindeutig identifiziert wird. Dies kann beispielsweise ein ID-Feld sein, wie *CustomerID*. IBM® SPSS® Modeler schreibt vor, dass dieses Schlüsselfeld zwingend numerisch sein muss.

*Anmerkung:* Dieses Feld ist für alle Oracle-Knoten mit Ausnahme von Oracle Adaptive Bayes, Oracle O-Cluster und Oracle Apriori optional.

**Automatische Datenvorbereitung.** (Nur 11g only) Aktiviert (Standard) oder deaktiviert den automatisierten Datenvorbereitungsmodus von Oracle Data Mining. Wenn dieses Kästchen markiert ist, führt ODM automatisch die vom Algorithmus geforderten Datenumformungen durch. Weitere Informationen finden Sie unter *Oracle Data Mining Concepts*.

**Maximale Regellänge.** Legt die maximale Anzahl an Vorbedingungen für jede beliebige Regel als ganze Zahl von 2 bis 20 fest. Auf diese Weise können Sie die Komplexität der Regeln begrenzen. Wenn Regeln zu komplex oder zu spezifisch sind oder das Training der Regelmenge zu viel Zeit in Anspruch nimmt, sollten Sie diese Einstellung reduzieren.

**Minimale Konfidenz.** Legt die minimale Konfidenzebene mit einem Wert zwischen 0 und 1 fest. Regeln mit einer niedrigeren Konfidenz als das angegebene Kriterium werden verworfen.

**Minimaler Support.** Legt die Untergrenze für Support als Wert zwischen 0 und 1 fest. A Priori erkennt Muster mit einer Häufigkeit über der Untergrenze für Support.

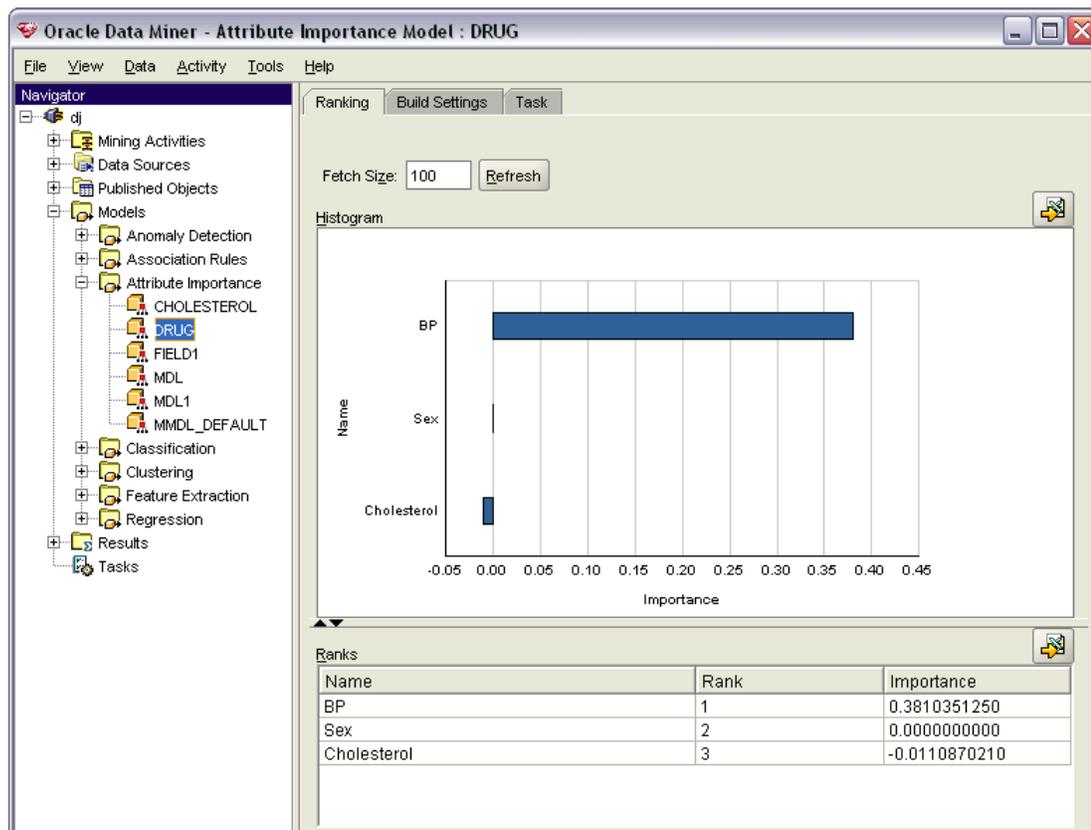
## Oracle Minimum Description Length (MDL)

Der Algorithmus "Oracle Minimum Description Length (MDL)" dient zur Ermittlung der Attribute, die den größten Einfluss auf ein Zielattribut haben. Wenn Sie wissen, welche Attribute den größten Einfluss haben, haben Sie häufig einen besseren Einblick in Ihre Geschäftstätigkeiten, können diese besser verwalten und einfacher Aktivitäten modellieren. Außerdem können diese Attribute die Datentypen anzeigen, durch deren Hinzufügung Sie Ihre Modelle erweitern können. MDL kann verwendet werden, um zum Beispiel die Prozessattribute zu ermitteln, die für die Prognose der Qualität eines hergestellten Bauteils, der mit Kundenabwanderung verbundenen Faktoren oder der Gene, die mit der größten Wahrscheinlichkeit in die Behandlung einer bestimmten Krankheit eingebunden sind, am relevantesten sind.

Oracle MDL verwirft Eingabefelder, die sie für die Vorhersage des Ziels als unwichtig erachtet. Mit den verbleibenden Eingabefeldern erstellt sie dann ein nicht verfeinertes Modell-Nugget, das mit einem Oracle-Modell verknüpft und in Oracle Data Miner sichtbar ist. Bei der Darstellung des Modells in Oracle Data Miner wird ein Diagramm angezeigt, das die übrigen Eingabefelder in der Reihenfolge ihrer Bedeutung zur Vorhersage des Ziels aufführt.

Abbildung 4-27

Verwenden des Oracle MDL-Diagramms zur Anzeige der relativen Wichtigkeit der Eingabefelder zur Vorhersage eines Ziels



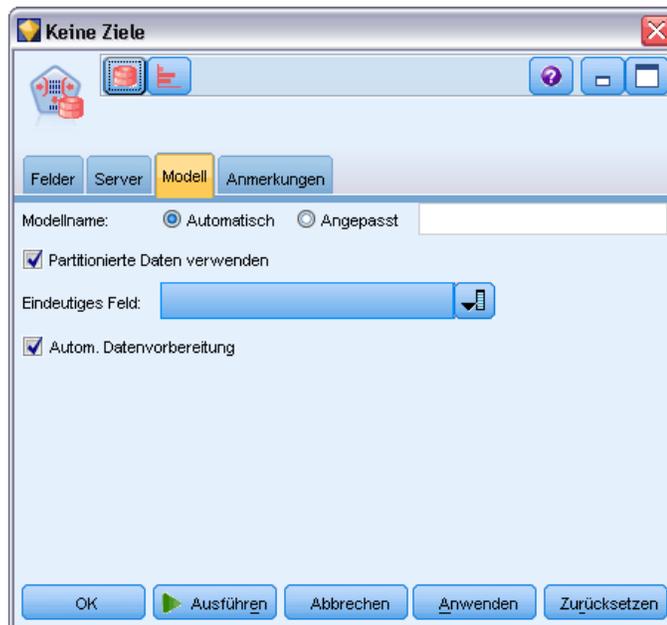
Negative Rangfolge bedeutet Rauschen. Eingabefelder, die bei null oder darunter eingeordnet werden, tragen nicht zur Vorhersage bei und sollten wahrscheinlich aus den Daten entfernt werden.

**So zeigen Sie das Diagramm an:**

- ▶ Klicken Sie mit der rechten Maustaste auf das nicht verfeinerte Modell-Nugget in der Modellpalette und wählen Sie Durchsuchen.
- ▶ Klicken Sie im Modellfenster auf die Schaltfläche zum Start von Oracle Data Miner.
- ▶ Bauen Sie eine Verbindung zu Oracle Data Miner auf. [Für weitere Informationen siehe Thema Oracle Data Miner auf S. 96.](#)
- ▶ Erweitern Sie im Oracle Data Miner-Navigationsfenster den Bereich Modelle und dann Attribut-Wichtigkeit.
- ▶ Wählen Sie das relevante Oracle-Modell aus (es hat denselben Namen wie in IBM® SPSS® Modeler angegebene Zielfeld). Wenn Sie nicht sicher sind, ob es das korrekte Modell ist, wählen Sie den Ordner “Attribut-Wichtigkeit” aus und suchen Sie ein Modell nach Erstellungsdatum.

## MDL-Modelloptionen

Abbildung 4-28  
MDL-Modelloptionen



**Modellname.** Sie können den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen (oder dem Modelltyp in Fällen, in denen kein solches Feld angegeben ist) generieren oder einen benutzerdefinierten Namen eingeben.

**Eindeutiges Feld.** Gibt das Feld an, über das jeder Fall eindeutig identifiziert wird. Dies kann beispielsweise ein ID-Feld sein, wie *CustomerID*. IBM® SPSS® Modeler schreibt vor, dass dieses Schlüsselfeld zwingend numerisch sein muss.

*Anmerkung:* Dieses Feld ist für alle Oracle-Knoten mit Ausnahme von Oracle Adaptive Bayes, Oracle O-Cluster und Oracle Apriori optional.

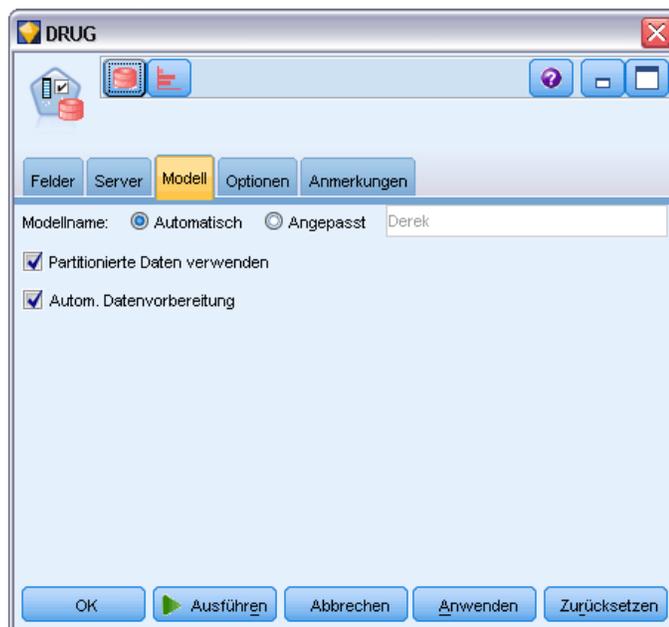
**Automatische Datenvorbereitung.** (Nur 11g only) Aktiviert (Standard) oder deaktiviert den automatisierten Datenvorbereitungsmodus von Oracle Data Mining. Wenn dieses Kästchen markiert ist, führt ODM automatisch die vom Algorithmus geforderten Datenumformungen durch. Weitere Informationen finden Sie unter *Oracle Data Mining Concepts*.

## Oracle Attribute Importance (AI)

Das Ziel der Attribut-Wichtigkeit ist es, die Attribute im Datenset zu finden, die mit dem Ergebnis zusammenhängen, sowie das Maß, in dem sie das Endergebnis beeinflussen. Der Knoten "Oracle Attribute Importance" analysiert Daten, findet Muster und sagt Ergebnisse mit einem entsprechenden Niveau an Zuverlässigkeit voraus.

## Alle Modelloptionen

Abbildung 4-29  
Alle Modelloptionen



**Modellname.** Sie können den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen (oder dem Modelltyp in Fällen, in denen kein solches Feld angegeben ist) generieren oder einen benutzerdefinierten Namen eingeben.

**Partitionierte Daten verwenden.** Wenn ein Partitionsfeld definiert ist, gewährleistet diese Option, dass nur Daten aus der Trainingspartition für die Modellerstellung verwendet werden. [Für weitere Informationen siehe Thema Partitionsknoten in Kapitel 4 in IBM SPSS Modeler 14.2- Quellen-, Prozess- und Ausgabeknoten.](#)

**Automatische Datenvorbereitung.** (Nur 11g only) Aktiviert (Standard) oder deaktiviert den automatisierten Datenvorbereitungsmodus von Oracle Data Mining. Wenn dieses Kästchen markiert ist, führt ODM automatisch die vom Algorithmus geforderten Datenumformungen durch. Weitere Informationen finden Sie unter *Oracle Data Mining Concepts*.

## Alle Auswahloptionen

Auf der Registerkarte “Optionen” können Sie die Standardeinstellungen für die Auswahl bzw. den Ausschluss von Eingabefeldern im Modell-Nugget angeben. Anschließend können Sie das Modell zu einem Stream hinzufügen, um die Untermenge der Felder auszuwählen, die in nachfolgenden Modellerstellungsvorgängen verwendet werden sollen. Alternativ können Sie diese Einstellungen nach der Modellgeneration durch die Auswahl bzw. das Aufheben der Auswahl weiterer Felder im Modellbrowser überschreiben. Die Standardeinstellungen ermöglichen es jedoch, das Modell-Nugget ohne weitere Änderungen anzuwenden, was insbesondere für die Skripterstellung nützlich sein kann.

Abbildung 4-30  
Alle Auswahloptionen



Die folgenden Optionen sind verfügbar:

**Alle Felder mit Rangzahl.** Wählt die Felder auf der Grundlage ihres Ranges (*bedeutsam*, *marginal* oder *unbedeutend*) aus. Sie können die Beschriftung für jeden Rang bearbeiten sowie die Cutoff-Werte ändern, die verwendet werden um Datensätze einem bestimmten Rang zuzuweisen.

**Obere Anzahl an Feldern.** Wählt die obersten  $n$  Felder nach Wichtigkeit aus.

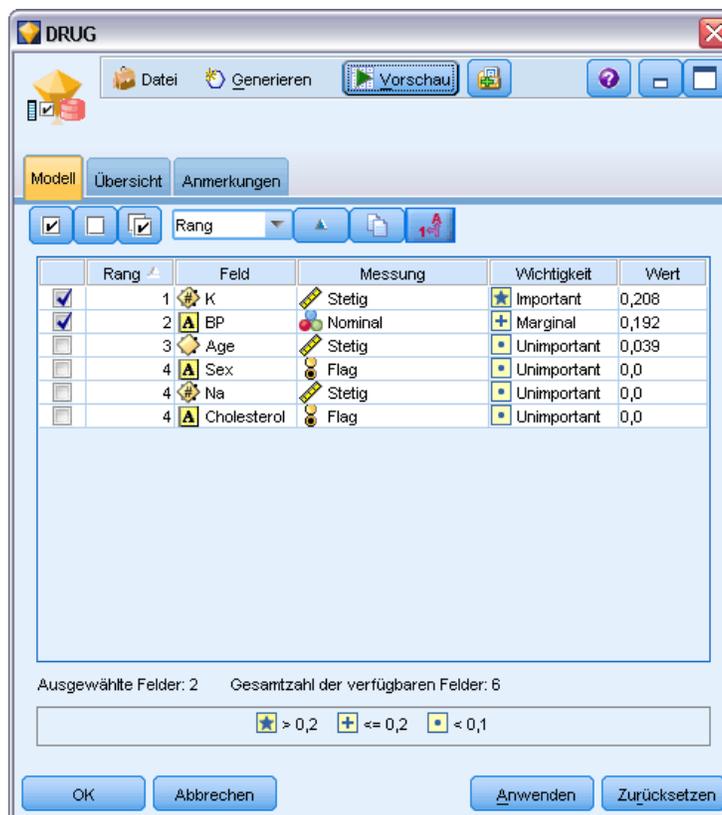
**Wichtigkeit größer als.** Wählt alle Felder aus, deren Wichtigkeit den angegebenen Wert übersteigt.

Das Zielfeld bleibt unabhängig von der Auswahl immer erhalten.

### AI-Modell-Nugget – Registerkarte “Modell”

Auf der Registerkarte “Modell” für ein Oracle AI-Modell-Nugget werden die Rangwertung und die Bedeutsamkeit für alle Eingaben angezeigt. Außerdem haben Sie die Möglichkeit, mithilfe der Kontrollkästchen in der Spalte auf der linken Seite Felder für die Filterung auszuwählen. Bei der Ausführung des Streams werden nur die aktivierten Felder sowie die Zielvorhersage beibehalten. Die anderen Eingabefelder werden verworfen. Die Standardauswahl beruht auf den im Modellierungsknoten angegebenen Optionen, Sie können jedoch nach Bedarf weitere Felder auswählen bzw. deren Auswahl aufheben.

Abbildung 4-31  
AI-Modell-Nugget



- Um die Liste nach Rang, Feldname, Wichtigkeit oder einer anderen der angezeigten Spalten zu sortieren, klicken Sie auf die Spaltenüberschrift. Wählen Sie alternativ das gewünschte Element neben der Schaltfläche “Sortieren nach” aus. Mit den nach unten bzw. oben zeigenden Pfeilen können Sie die Sortierrichtung ändern.

- Sie können mithilfe der Symbolleiste alle Felder aktivieren bzw. deaktivieren und auf das Dialogfeld “Felder markieren” zugreifen, in dem Sie Felder nach Rangordnung oder Wichtigkeit auswählen können. Zum Erweitern der Auswahl können Sie auch die Umschalt- oder Strg-Taste drücken, während Sie auf Felder klicken. [Für weitere Informationen siehe Thema Auswählen der Felder nach Wichtigkeit in Kapitel 4 in IBM SPSS Modeler 14.2-Modellierungsknoten.](#)
- Die Schwellenwerte für die Einordnung von Eingaben als “bedeutsam”, “marginal” bzw. “unbedeutend” werden in der Legende unterhalb der Tabelle angezeigt. Diese Werte werden im Modellierungsknoten angegeben.

## **Verwalten von Oracle-Modellen**

Oracle-Modelle werden genau wie andere IBM® SPSS® Modeler-Modelle zur Modellpalette hinzugefügt und können fast genauso benutzt werden. Es gibt jedoch einige wichtige Unterschiede, die sich daraus ergeben, dass zurzeit jedes in SPSS Modeler erstellte Oracle-Modell auf ein in einem Datenbank-Server gespeichertes Modell verweist.

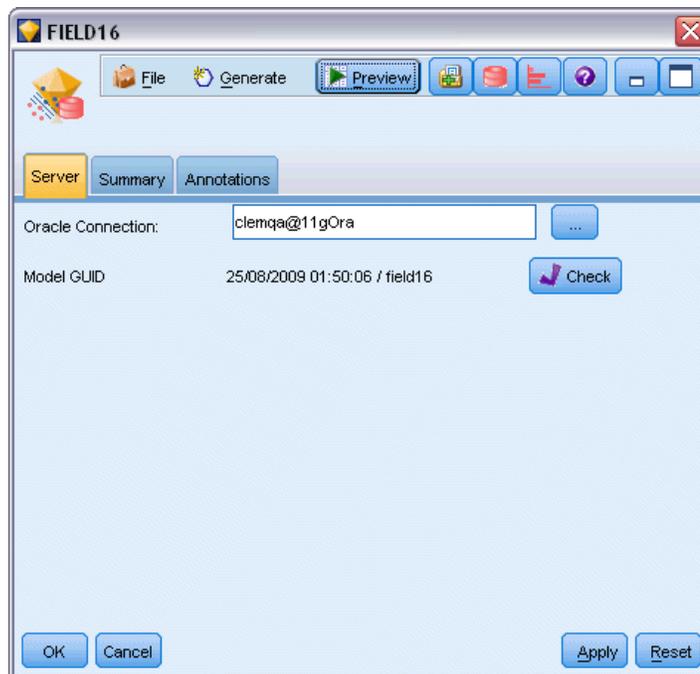
### **Oracle-Modell-Nugget – Registerkarte “Server”**

Bei der Bildung eines ODM-Modells mit IBM® SPSS® Modeler wird einerseits in SPSS Modeler ein Modell erzeugt und andererseits in der Oracle-Datenbank ein Modell erzeugt oder ersetzt. Ein derartiges SPSS Modeler-Modell stellt einen Bezug zum Inhalt eines in einem Datenbankserver gespeicherten Datenbankmodells her. SPSS Modeler kann eine Konsistenzprüfung durchführen, indem sowohl im SPSS Modeler-Modell als auch im Oracle-Modell eine identische, generierte Zeichenkette vom Typ **Modellschlüssel** gespeichert wird.

Die Schlüsselzeichenkette wird für jedes Oracle-Modell im Dialogfeld “Modelle auflisten” in der Spalte *Modellinformationen* angezeigt. Die Schlüsselzeichenkette eines SPSS Modeler-Modells wird auf der Registerkarte “Server” eines SPSS Modeler-Modells (wenn es sich in einem Stream befindet) als Modellschlüssel ausgegeben.

Mit der Schaltfläche “Überprüfen” auf der Registerkarte “Server” eines Modell-Nuggets wird ermittelt, ob die Modellschlüssel des SPSS Modeler-Modells und des Oracle-Modells übereinstimmen. Wenn in Oracle kein Modell mit demselben Namen gefunden wird oder wenn der Modellschlüssel nicht übereinstimmt, bedeutet dies, dass das Oracle-Modell seit der Bildung des SPSS Modeler-Modells gelöscht oder neu erstellt wurde.

Abbildung 4-32  
Oracle-Modell-Nugget – Optionen der Registerkarte “Server”



### Oracle-Modell-Nugget – Registerkarte “Übersicht”

Auf der Registerkarte “Übersicht” eines Modell-Nuggets finden Sie Informationen zum Modell selbst (*Analyse*), den im Modell verwendeten Feldern (*Felder*), den beim Erstellen des Modells verwendeten Einstellungen (*Aufbaueinstellungen*) und zum Modelltraining (*Trainingsübersicht*).

Beim ersten Durchsuchen des Knotens sind die Ergebnisse der Registerkarte “Übersicht” reduziert. Um die für Sie relevanten Ergebnisse anzuzeigen, vergrößern Sie sie mithilfe des Erweiterungssteuerelements links neben dem betreffenden Element oder klicken Sie auf die Schaltfläche Alles anzeigen, um alle Ergebnisse anzuzeigen. Um die Ergebnisse nach der Betrachtung wieder auszublenden, können Sie mit dem Erweiterungssteuerelement die gewünschten Ergebnisse reduzieren. Alternativ können Sie mit der Schaltfläche Alles ausblenden alle Ergebnisse ausblenden.

**Analyse.** Zeigt Informationen zum jeweiligen Modell an. Wenn Sie einen Analyseknoden ausgeführt haben, der an dieses Modell-Nugget angehängt ist, werden die Informationen aus dieser Analyse ebenfalls in diesem Abschnitt angezeigt. [Für weitere Informationen siehe Thema Analyseknoden in Kapitel 6 in IBM SPSS Modeler 14.2- Quellen-, Prozess- und Ausgabeknoten.](#)

**Felder.** Listet die Felder auf, die bei der Erstellung des Modells als Ziel bzw. als Eingaben verwendet werden.

**Aufbaueinstellungen.** Enthält Informationen zu den beim Aufbau des Modells verwendeten Einstellungen.

**Trainingsübersicht.** In diesem Abschnitt werden folgende Informationen angezeigt: der Typ des Modells, der für seine Erstellung verwendete Stream, der Benutzer, der ihn erstellt hat, der Erstellungszeitpunkt und die für den Aufbau des Modells benötigte Zeit.

### ***Oracle-Modell-Nugget – Registerkarte “Einstellungen”***

In der Registerkarte “Einstellungen” des Modell-Nugget können Sie die Einstellung bestimmter Optionen für den Modellierungsknoten zu Scoring-Zwecken überschreiben.

#### ***Oracle Decision Tree***

**Fehlklassifizierungskosten verwenden.** Bestimmt, ob Fehlklassifizierungskosten im Oracle Decision Tree-Modell verwendet werden. [Für weitere Informationen siehe Thema Fehlklassifizierungskosten auf S. 60.](#)

**Regel-ID.** Wenn ausgewählt (markiert), wird dem Oracle Decision Tree-Modell eine Regel-ID-Spalte hinzugefügt. Die Regel-ID identifiziert den Knoten in der Struktur, an dem eine bestimmte Aufteilung erfolgt.

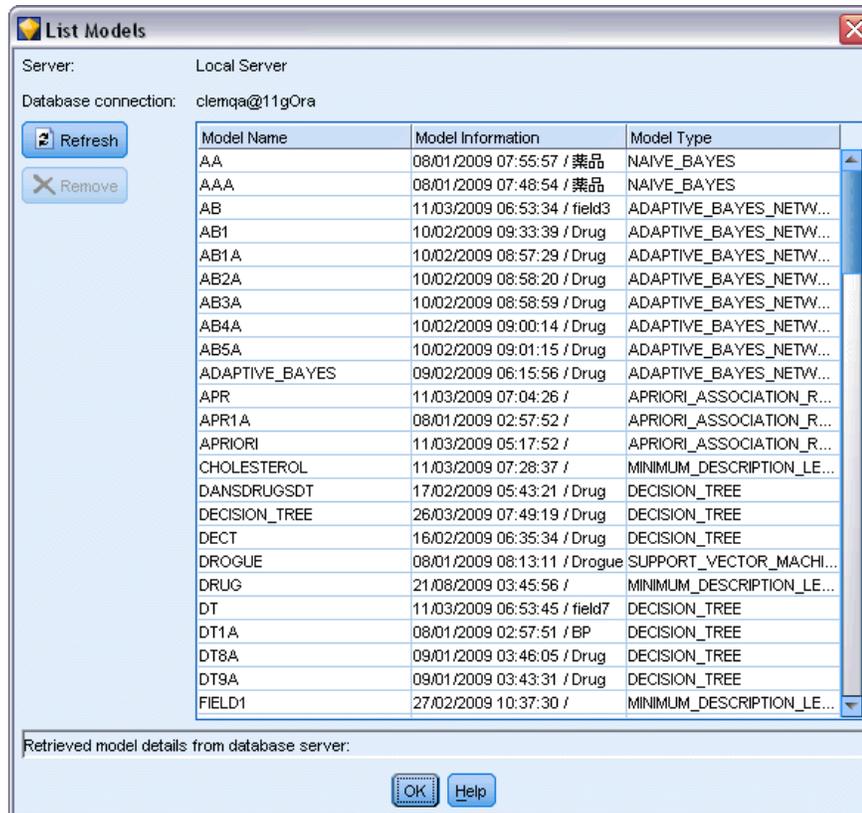
#### ***Oracle NMF***

**Alle Merkmale anzeigen.** Wenn ausgewählt (markiert), wird die Merkmals-ID und die Konfidenz für alle Merkmale im Oracle NMF-Modell angezeigt und nicht nur die Werte für das beste Merkmal.

### ***Auflisten der Oracle-Modelle***

Die Schaltfläche “Oracle Data Mining-Modelle auflisten” öffnet ein Dialogfeld, in dem die vorhandenen Datenbankmodelle aufgelistet sind und entfernt werden können. Der Zugriff auf dieses Dialogfeld erfolgt über das Dialogfeld “Hilfsprogramme” und über die Dialogfelder zum Erstellen, Suchen und Anwenden für mit ODM verbundene Knoten.

Abbildung 4-33  
Dialogfeld "Oracle List Models"



Zu jedem Modell werden folgende Informationen angezeigt:

- **Modellname.** Name des Modells, das zum Sortieren der Liste verwendet wird
- **Modellinformationen.** Modellschlüsselinformationen, die sich aus dem Datum und der Uhrzeit der Erstellung sowie dem Namen der Zielspalte zusammensetzen
- **Modelltyp.** Name des Algorithmus, der dieses Modell erstellt hat

## Oracle Data Miner

Oracle Data Miner ist die Benutzeroberfläche für Oracle Data Mining (ODM) und ersetzt die frühere IBM® SPSS® Modeler-Benutzeroberfläche für ODM. Oracle Data Miner dient zur Erhöhung der Erfolgsquote des Analysten bei der ordnungsgemäßen Nutzung der ODM-Algorithmen. Es werden verschiedene Methoden verwendet, um diese Ziele zu erreichen:

- Die Benutzer benötigen mehr Unterstützung bei der Anwendung einer Methodologie, die sich sowohl mit der Datenvorbereitung als auch mit der Algorithmenauswahl befasst. Oracle Data Miner wird dem durch Bereitstellung von Data-Mining-Aktivitäten gerecht, mit dem die Benutzer Schritt für Schritt durch die jeweilige Methodologie geführt werden.
- Oracle Data Miner beinhaltet verbesserte und erweiterte Heuristiken in den Modellerstellungs- und Transformationsassistenten, um die Fehlerwahrscheinlichkeit bei der Angabe von Modell- und Transformationseinstellungen zu verringern.

### Definieren einer Oracle Data Miner-Verbindung

- ▶ Oracle Data Miner kann von allen Oracle-Dialogfeldern für Erstellung, Knotenanwendung und Ausgabe mithilfe der Schaltfläche **Oracle Data Miner starten** gestartet werden.

Abbildung 4-34  
Schaltfläche "Oracle Data Miner starten"



- ▶ Das Oracle Data Miner-Dialogfeld **Edit Connection** wird dem Benutzer angezeigt, bevor die externe Oracle Data Miner-Anwendung gestartet wird (vorausgesetzt, die Option für Hilfsprogramme wurde ordnungsgemäß definiert).

*Hinweis:* Dieses Dialogfeld wird nur angezeigt, wenn kein definierter Verbindungsname vorhanden ist.

Abbildung 4-35  
Oracle Data Miner – Dialogfeld "Edit Connection"

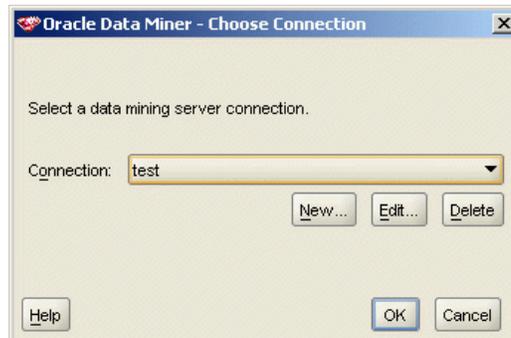


- Geben Sie einen Data Miner-Verbindungsnamen und die entsprechenden Informationen für den Oracle 10gR1- bzw. 10gR2-Server ein. Bei dem Oracle-Server sollte es sich um denselben Server handeln, der auch in SPSS Modeler angegeben wurde.

- Das Dialogfeld **Choose Connection** des Oracle Data Miner bietet Optionen, in denen Sie angeben können, welcher Verbindungsname (im Schritt weiter oben definiert), verwendet werden soll.

Abbildung 4-36

Oracle Data Miner – Dialogfeld "Choose Connection"

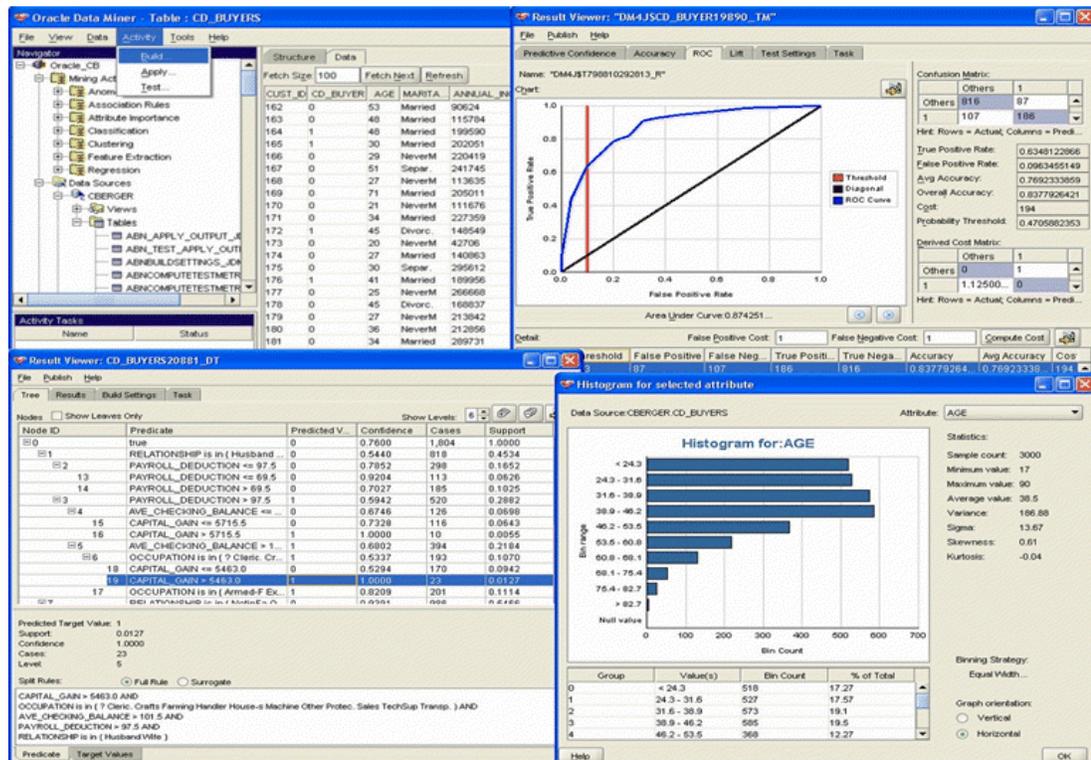


Unter [Oracle Data Miner](http://www.oracle.com/technology/products/bi/odm/odminer/odminer_install_102.htm)

([http://www.oracle.com/technology/products/bi/odm/odminer/odminer\\_install\\_102.htm](http://www.oracle.com/technology/products/bi/odm/odminer/odminer_install_102.htm)) auf der Oracle Website finden Sie weitere Informationen zu Anforderungen, Installation und Verwendung von Oracle Data Miner.

Abbildung 4-37

Benutzeroberfläche von Oracle Data Miner



## Vorbereitung der Daten

Wenn Sie bei der Modellbildung die mit den Oracle Data Mining gelieferten Algorithmen Naive Bayes, Adaptive Bayes und Support Vector Machine nutzen, können sich zwei Arten der Datenvorbereitung als nützlich erweisen:

- **Klassieren** oder die Konvertierung fortlaufender numerischer Bereichsfelder in Kategorien für Algorithmen, die keine fortlaufenden Daten annehmen.
- **Normalisierung** oder auf numerischen Bereichen durchgeführte Transformationen, die dafür sorgen, dass diese ähnliche Bedeutungen und Standardabweichungen besitzen.

### **Klassierung**

Der Binning-Knoten von IBM® SPSS® Modeler bietet diverse Techniken für Binning-Operationen. Eine Klassieroperation ist definiert und kann auf eines oder mehrere Felder angewendet werden. Die Ausführung der Klassieroperation auf einem Daten-Set erzeugt die Grenzwerte und ermöglicht das Erstellen eines SPSS Modeler-Ableitungsknotens. Die Ableitungsoperation kann in SQL konvertiert werden und vor der Bildung und dem Scoring des Modells angewendet werden. Dieser Ansatz erzeugt eine Abhängigkeit zwischen dem Modell und dem die Klassierung durchführenden Ableitungsknoten, erlaubt aber, dass die Klassierspezifikationen von verschiedenen Modellbildungsaufgaben wiederverwendet werden.

### **Normalisierung**

Stetige Felder (numerischer Bereich), die als Eingabe für SVM-Modelle verwendet werden, sollten vor der Modellbildung normalisiert werden. Bei einem Regressionsmodell muss die Normalisierung außerdem umgekehrt werden, um den Score der Modellausgabe zu rekonstruieren. Als SVM-Modelleinstellungen stehen Z-Score, Min-Max oder Keine zur Auswahl. Die Normalisierungskoeffizienten werden von Oracle im Rahmen des Modellbildungsvorgangs erzeugt und dann in SPSS Modeler geladen, wo sie zusammen mit dem Modell gespeichert werden. Zum Zeitpunkt der Anwendung werden die Koeffizienten in SPSS Modeler-Ableitungsausdrücke konvertiert und zur Vorbereitung der Daten für das Scoring verwendet, bevor diese an das Modell übergeben werden. In diesem Fall ist die Normalisierung eng mit der Modellbildungsaufgabe verbunden.

## Beispiele für Oracle Data Mining

Im Lieferumfang von Clementine sind einige Beispiel-Streams enthalten, die die Verwendung von ODM mit IBM® SPSS® Modeler demonstrieren. Diese Streams befinden sich im SPSS Modeler-Installationsverzeichnis unter `\Demos\Database_Modelling\Oracle Data Mining\`.

*Hinweis:* Dieser Demo-Ordner kann über die Programmgruppe "SPSS Modeler" im Windows-Startmenü aufgerufen werden.

Die folgenden Streams können als Abfolge gemeinsam als Beispiel für einen Database-Mining-Prozess verwendet werden, bei dem der SVM-Algorithmus (Support Vector Machine) von Oracle Data Mining genutzt wird.

Stream	Beschreibung
<i>1_upload_data.str</i>	Bereinigt Daten und lädt sie aus einer Textdatei in die Datenbank.
<i>2_explore_data.str</i>	Bietet ein Beispiel für die Datenuntersuchung mit SPSS Modeler.
<i>3_build_model.str</i>	Erstellt das Modell unter Verwendung des datenbankeigenen Algorithmus.
<i>4_evaluate_model.str</i>	Wird als Beispiel für die Modellevaluation mit SPSS Modeler verwendet.
<i>5_deploy_model.str</i>	Verwendet das Modell für datenbankinternes Scoring.

*Anmerkung:* Um das Beispiel auszuführen, müssen die Streams in der richtigen Reihenfolge ausgeführt werden. Außerdem müssen die Quellen- und Modellierungsknoten in den einzelnen Streams aktualisiert werden, um auf eine gültige Datenquelle für die zu verwendende Datenbank zu verweisen.

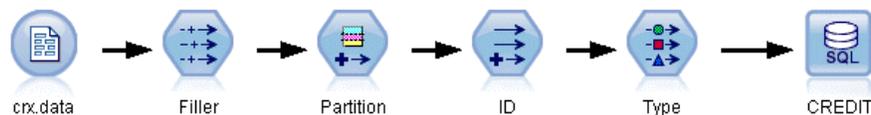
Die in den Beispiel-Streams verwendeten Datenmengen beziehen sich auf Kreditkartenanwendungen und stellen ein Klassifizierungsproblem mit einer Mischung aus kategorialen und stetigen Prädiktoren dar. Weitere Informationen über diese Datenmenge finden Sie in der Datei *crx.names* im selben Ordner wie die Beispiel-Streams.

Diese Daten stehen im UCI Machine Learning Repository unter der folgenden Adresse zur Verfügung: <ftp://ftp.ics.uci.edu/pub/machine-learning-databases/credit-screening/>.

### Beispiel-Stream: Hochladen von Daten

Der erste Beispiel-Stream, *1\_upload\_data.str*, wird verwendet, um Daten aus einer Textdatei zu bereinigen und in Oracle zu laden.

Abbildung 4-38  
Beispiel-Stream zum Hochladen von Daten



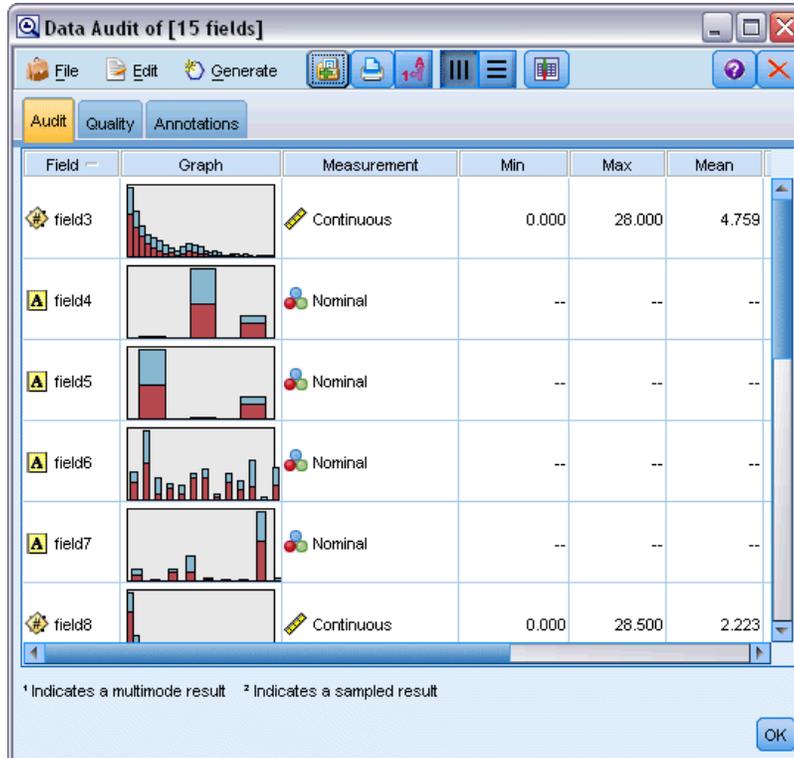
Da für Oracle Data Mining ein eindeutiges ID-Feld erforderlich ist, fügt dieser erste Stream mithilfe eines Ableitungsknotens mit der @INDEX-Funktion von IBM® SPSS® Modeler dem Daten-Set ein neues Feld mit dem Namen *ID* und den eindeutigen Werten "1,2,3" hinzu.

Der Füllerknoten ist für die Behandlung von fehlenden Werten zuständig und ersetzt leere, aus der Textdatei *crx.data* eingelesene Felder durch *NULL*-Werte.

### Beispiel-Stream: Untersuchen von Daten

Der zweite Beispiel-Stream, *2\_explore\_data.str*, soll zeigen, wie mithilfe eines Data Audit-Knotens ein allgemeiner Überblick über die Daten (einschließlich statistischer Funktionen und Diagramme) gewonnen werden kann. Für weitere Informationen siehe Thema [Data Audit-Knoten in Kapitel 6 in IBM SPSS Modeler 14.2- Quellen-, Prozess- und Ausgabeknoten](#).

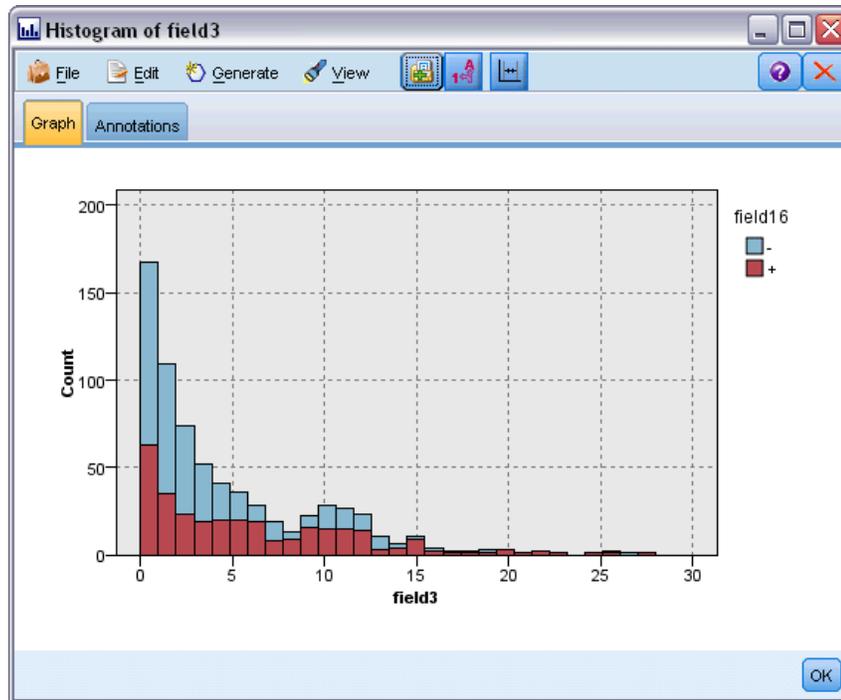
Abbildung 4-39  
Data Audit-Ergebnisse



Wenn Sie im Data Audit-Bericht auf ein Diagramm doppelklicken, wird ein detaillierteres Diagramm angezeigt, in dem Sie einzelne Felder eingehender untersuchen können.

Abbildung 4-40

Im Fenster "Data Audit" durch Doppelklicken auf dem Diagramm erzeugtes Histogramm

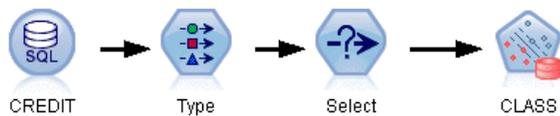


### Beispiel-Stream: Erstellen des Modells

Der dritte Beispiel-Stream, *3\_build\_model.str*, veranschaulicht die Modellerstellung in IBM® SPSS® Modeler. Doppelklicken Sie auf den Datenbank-Quellenknoten (mit der Beschriftung CREDIT), um die Datenquelle anzugeben. Um die Einstellungen für die Erstellung festzulegen, doppelklicken Sie auf den Erstellungsknoten (ursprünglich mit der Beschriftung "CLASS", die sich in "FIELD16" ändert, wenn die Datenquelle angegeben wurde).

Abbildung 4-41

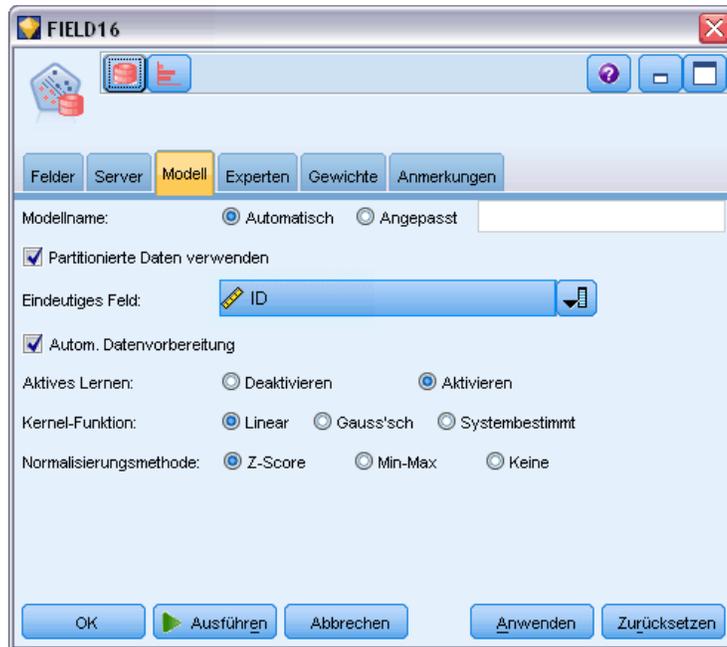
Beispiel-Stream für die Datenbank-Modellbildung



Gehen Sie auf der Registerkarte "Modell" des Dialogfelds wie folgt vor:

- ▶ Vergewissern Sie sich, dass ID als eindeutiges Feld ausgewählt ist.
- ▶ Vergewissern Sie sich, dass als Kernel-Funktion Linear und als Normalisierungsmethode Z-Score ausgewählt ist.

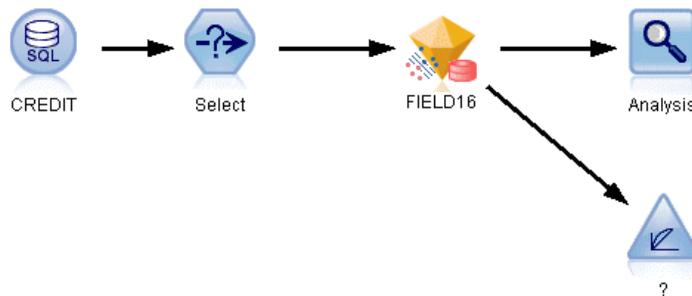
Abbildung 4-42  
 Modelloptionen für Oracle SVM



### Beispiel-Stream: Evaluieren des Modells

Der vierte Beispiel-Stream, *4\_evaluate\_model.str*, veranschaulicht die Vorteile der Verwendung von IBM® SPSS® Modeller für die datenbankinterne Modellbildung. Sobald Sie das Modell ausgeführt haben, können Sie es wieder zu Ihrem Daten-Stream hinzufügen und das Modell mit verschiedenen von SPSS Modeller bereitgestellten Tools evaluieren.

Abbildung 4-43  
 Beispiel-Stream für die Modellevaluation



### Anzeigen der Modellbildungsergebnisse

Gliedern Sie einen Tabellenknoten an das Modell-Nugget an, um die Ergebnisse zu untersuchen. Das Feld \$O-field16 zeigt den vorhergesagten Wert für *field16* für jeden Fall, und das Feld \$OC-field16 zeigt den Konfidenzwert für diese Vorhersage.

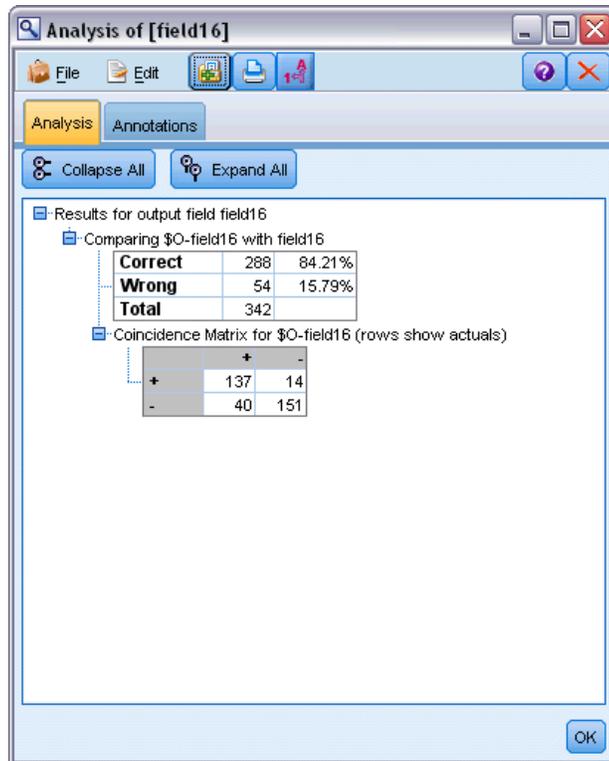
Abbildung 4-44  
Tabelle mit Informationen zu generierten Vorhersagen

	field12	field13	field14	field15	field16	Partition	ID	\$O-field16	\$OC-field16
1		g	300	0	-	2_Test...	454	-	0.818
2		g	320	3552	-	2_Test...	456	-	0.818
3		g	240	0	-	2_Test...	458	-	0.820
4		g	160	0	-	2_Test...	460	-	0.819
5		g	360	0	-	2_Test...	463	-	0.819
6		g	200	18	-	2_Test...	464	-	0.820
7		g	320	5	-	2_Test...	471	-	0.820
8		g	360	1000	-	2_Test...	474	-	0.819
9		g	220	5	-	2_Test...	477	-	0.819
10		s	80	0	-	2_Test...	480	-	0.819
11		g	240	35	-	2_Test...	481	-	0.817
12		g	280	80	-	2_Test...	482	-	0.819
13		g	128	6	-	2_Test...	484	-	0.819
14		g	0	351	-	2_Test...	486	-	0.822
15		g	180	1	-	2_Test...	489	-	0.822
16		g	333	892	+	2_Test...	491	+	0.818
17		g	520	2000	+	2_Test...	492	+	0.819
18		g	340	0	+	2_Test...	494	+	0.817
19		g	240	0	+	2_Test...	495	+	0.816
20		g	160	5860	+	2_Test...	497	+	0.819

### **Evaluieren der Modellergebnisse**

Sie können den Analyseknoden verwenden, um eine Fehlklassifizierungstabelle zu erstellen, aus der das Muster der Übereinstimmungen zwischen jedem vorhergesagten Feld und dem zugehörigen Zielfeld ersichtlich wird. Führen Sie den Analyseknoden aus, um die Ergebnisse anzuzeigen.

Abbildung 4-45  
Registerkarte "Analyse" mit Informationen zu den Analyseergebnissen

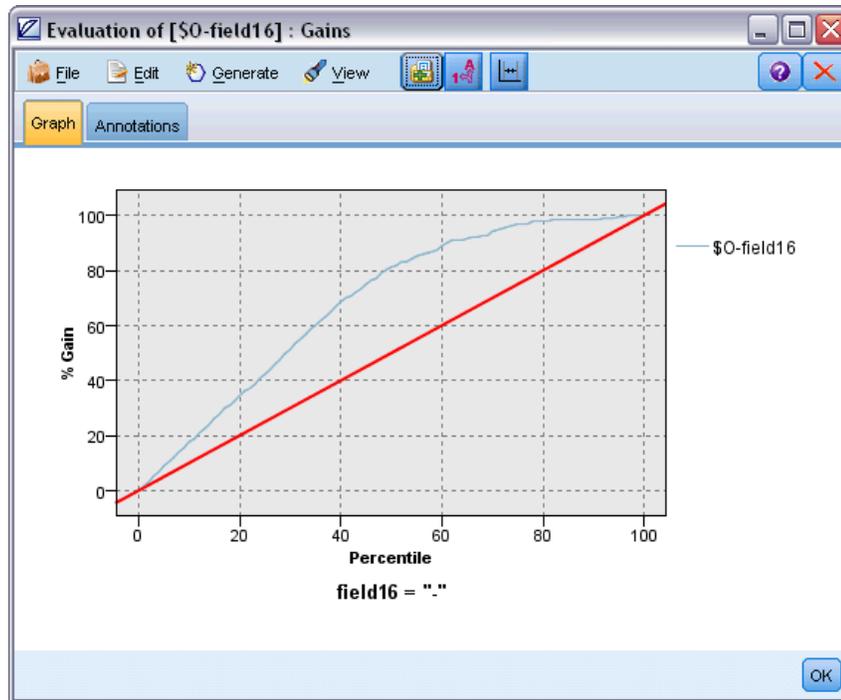


Aus der Tabelle geht hervor, dass 84,21 % der vom Oracle SVM-Algorithmus erstellten Vorhersagen richtig sind.

Sie können mithilfe des Evaluationsknotens ein Gewinnendiagramm erstellen, das die Verbesserungen der Vorhersagegenauigkeit durch das Modell aufzeigt. Führen Sie den Evaluationsknoten aus, um die Ergebnisse anzuzeigen.

Abbildung 4-46

Gewinndiagramm mit Informationen zur Verbesserung der Vorhersagegenauigkeit für das Modell

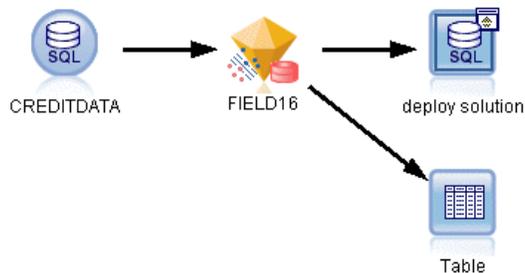


### Beispiel-Stream: Bereitstellen des Modells

Sobald Sie mit der Genauigkeit des Modells zufrieden sind, können Sie es für die Verwendung mit externen Anwendungen oder für eine erneute Veröffentlichung in der Datenbank bereitstellen. Im letzten Beispiel-Stream, *5\_deploy\_model.str*, werden Daten aus der Tabelle **CREDITDATA** gelesen und dann mit dem Publisher-Knoten *deploy solution* gesort und in der Tabelle **CREDITSCORES** veröffentlicht.

Abbildung 4-47

Beispiel-Stream für die Datenbank-Modellbildung



Für weitere Informationen siehe Thema [So funktioniert IBM SPSS Modeler Solution Publisher](#) in Kapitel 2 in *IBM SPSS Modeler 14.2 Solution Publisher*.

# **Datenbankmodellierung mit IBM InfoSphere Warehouse**

## **IBM InfoSphere Warehouse und IBM SPSS Modeler**

IBM InfoSphere Warehouse (ISW) bietet eine Serie von in IBM DB2 RDBMS integrierten Data-Mining-Algorithmen. IBM® SPSS® Modeler bietet Knoten, die die Integration folgender IBM-Algorithmen unterstützen:

- Decision Trees (Entscheidungsbäume)
- Assoziationsregeln
- Demografische Clusterbildung
- Kohonen-Clustering
- Sequenzregeln
- Transformationsregression
- Lineare Regression
- Polynomiale Regression
- Naive Bayes
- Logistische Regression
- Zeitreihen

Weitere Informationen über diese Algorithmen finden Sie in der Begleitdokumentation zu Ihrer IBM InfoSphere Warehouse-Installation.

## **Anforderungen für die Integration mit InfoSphere Warehouse**

Für die datenbankinterne Modellbildung mit InfoSphere Warehouse Data Mining gelten die folgenden Voraussetzungen. Wenden Sie sich ggf. an Ihren Datenbankverwalter, um sicherzustellen, dass diese Bedingungen erfüllt sind.

- IBM® SPSS® Modeler wird im Rahmen einer IBM® SPSS® Modeler Server-Installation unter Windows oder UNIX ausgeführt.
- IBM DB2 Data Warehouse Edition Version 9.1  
*oder*
- IBM InfoSphere Warehouse Version 9.5 Enterprise Edition
- Eine ODBC-Datenquelle für die Verbindung mit DB2, wie unten beschrieben.

*Hinweis:* Für Datenbankmodellierung und SQL-Optimierung muss auf dem SPSS Modeler-Computer die SPSS Modeler Server-Konnektivität aktiviert sein. Wenn diese Einstellung aktiviert ist, können Sie auf Datenbankalgorithmen zugreifen, SQL direkt aus SPSS Modeler per

Pushback übertragen und auf SPSS Modeler Server zugreifen. Um den aktuellen Lizenzstatus zu überprüfen, klicken Sie im Menü "Hilfe" in SPSS Modeler auf Info. [Für weitere Informationen siehe Thema Verbindung mit IBM SPSS Modeler Server in Kapitel 3 in IBM SPSS Modeler 14.2- Benutzerhandbuch.](#)

## **Aktivieren der Integration mit InfoSphere Warehouse**

Um die Integration von IBM® SPSS® Modeler mit IBM InfoSphere Warehouse (ISW) Data Mining zu ermöglichen, müssen Sie IBM InfoSphere Warehouse konfigurieren und eine ODBC-Datenquelle erstellen, im SPSS Modeler-Dialogfeld "Hilfsprogramme" die Integration aktivieren und die SQL-Erzeugung und -Optimierung aktivieren.

### **Konfigurieren von ISW**

Befolgen Sie zum Installieren und Konfigurieren von IWS die Anweisungen im Handbuch zur *InfoSphere Warehouse-Installation*.

### **Erstellen einer ODBC-Datenquelle für ISW**

Um die Verbindung zwischen ISW und SPSS Modeler zu aktivieren, müssen Sie einen ODBC-Datenquellennamen (DSN) erstellen.

Bevor Sie einen DSN erstellen, sollten Sie grundlegende Kenntnisse über ODBC-Datenquellen und -Treiber sowie über Datenbankunterstützung in SPSS Modeler besitzen. [Für weitere Informationen siehe Thema Datenzugriff in Kapitel 2 in IBM SPSS Modeler Server 14.2-Verwaltungs- und Leistungshandbuch.](#)

Wenn IBM® SPSS® Modeler Server und InfoSphere Warehouse Data Mining auf unterschiedlichen Hosts ausgeführt werden, müssen Sie auf beiden Hosts den gleichen ODBC-DSN erstellen. Stellen Sie sicher, dass Sie auf jedem Host denselben Namen für diesen DSN verwenden.

- ▶ Installieren Sie die ODBC-Treiber. Diese Treiber finden Sie auf dem zu dieser Version gehörenden IBM® SPSS® Data Access Pack-Installationsmedium. Führen Sie die Datei *setup.exe* aus, um das Installationsprogramm zu starten und wählen Sie alle relevanten Treiber aus. Folgen Sie den Anweisungen am Bildschirm, um die Treiber zu installieren.

- ▶ Erstellen Sie den DSN.

*Anmerkung:* Die Befehlsfolge ist abhängig von der jeweiligen Windows-Version.

- **Windows XP.** Wählen Sie im Menü "Start" die Option Systemsteuerung. Doppelklicken Sie auf Verwaltung und dann auf Datenquellen (ODBC).
  - **Windows Vista** Wählen Sie im Menü "Start" die Option Systemsteuerung und dann Systemwartung. Doppelklicken Sie auf Verwaltung, wählen Sie dann Datenquellen (ODBC) und klicken Sie auf Öffnen.
  - **Windows 7.** Wählen Sie im Menü "Start" die Option Systemsteuerung, dann System & Sicherheit und dann Verwaltung. Wählen Sie Datenquellen (ODBC) und klicken Sie dann auf Öffnen.
- ▶ Klicken Sie auf die Registerkarte System-DSN und dann auf Hinzufügen.

- ▶ Wählen Sie den Treiber SPSS OEM 6.0 DB2 Wire Protocol aus.
- ▶ Klicken Sie auf Fertig stellen.
- ▶ Gehen Sie im Dialogfeld zum Einrichten des ODBC DB2 Wire Protocol-Treibers wie folgt vor:
  - Geben Sie den Namen einer Datenquelle ein.
  - Geben Sie für die IP-Adresse den Hostnamen des Servers ein, auf dem sich DB2 RDBMS befindet.
  - Übernehmen Sie für den TCP-Port den Standardwert (50000).
  - Geben Sie den Namen der Datenbank an, mit der Sie eine Verbindung herstellen möchten.
- ▶ Klicken Sie auf Test Connection (Verbindung testen).
- ▶ Geben Sie im Dialogfeld “Logon to DB2 Wire Protocol” (Anmeldung bei DB2 Wire Protocol) den Benutzernamen und das Passwort ein, die Sie vom Datenbankadministrator erhalten haben, und klicken Sie auf OK.

Die Meldung Connection established! (Verbindung hergestellt!) wird angezeigt.

**IBM DB2 ODBC DRIVER.** Wenn Sie den IBM DB2 ODBC DRIVER als ODBC-Treiber einsetzen, erstellen Sie mit folgenden Schritten ein ODBC-DSN:

- ▶ Klicken Sie im der ODBC-Datenquellen-Administrator auf die Registerkarte System-DSN und dann auf Hinzufügen.
- ▶ Wählen Sie IBM DB2 ODBC DRIVER und klicken Sie auf Fertig stellen.
- ▶ Geben Sie im Fenster “Add” (Hinzufügen) für den IBM DB2 ODBC DRIVER den Namen einer Datenquelle an und klicken Sie für den Datenbank-Alias auf Add (Hinzufügen).
- ▶ Geben Sie im Fenster “<Name der Datenquelle>” unter “CLI/ODBC Settings” (CLI/ODBC-Einstellungen) auf der Registerkarte “Data Source” (Datenquelle) die Benutzer-ID und das Passwort ein, die Sie vom Datenbankadministrator erhalten haben, und klicken Sie anschließend auf die Registerkarte TCP/IP.
- ▶ Nehmen Sie auf der Registerkarte “TCP/IP” folgende Eingaben vor:
  - Den Namen der Datenbank, mit der Sie eine Verbindung herstellen möchten.
  - Einen Datenbank-Aliasnamen (maximal acht Zeichen).
  - Den Hostnamen des Datenbankservers, mit dem Sie eine Verbindung herstellen wollen.
  - Die Portnummer für die Verbindung.
- ▶ Klicken Sie auf die Registerkarte Security Options (Sicherheitsoptionen), wählen Sie Specify the security options (Optional) (Sicherheitsoptionen angeben (optional)) und übernehmen Sie die Standardeinstellung (Use authentication value in server’s DBM Configuration (Authentifizierungswert in der DBM-Konfiguration des Servers verwenden)).
- ▶ Klicken Sie auf die Registerkarte Data Source (Datenquelle) und anschließend auf Connect (Verbinden).

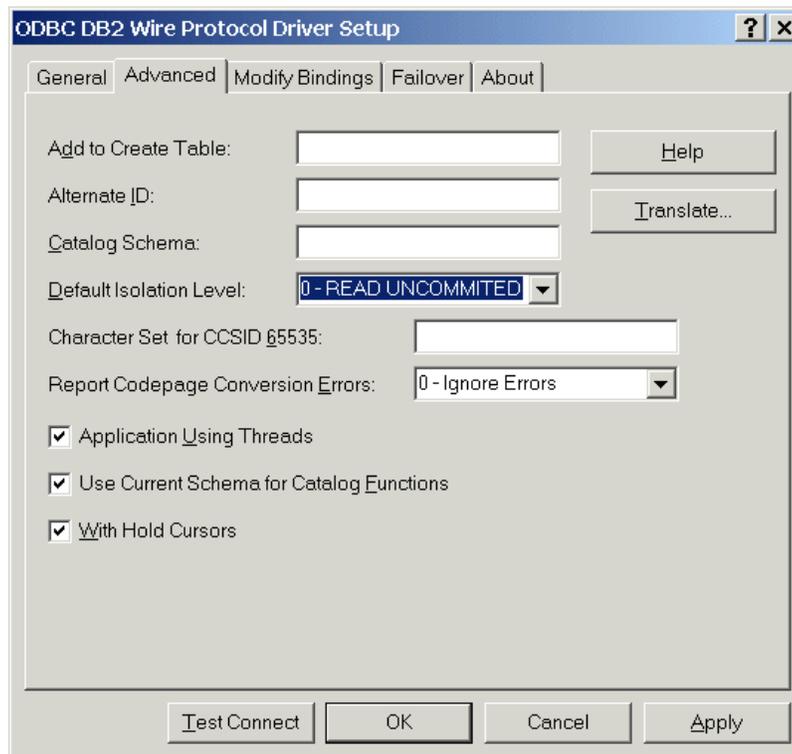
Die Meldung Connection tested successfully (Verbindung erfolgreich getestet) wird angezeigt.

**ODBC für Feedback konfigurieren (optional)**

Mit den folgenden Schritten konfigurieren Sie die im vorangegangenen Abschnitt erstellte ODBC-Datenquelle so, dass Sie von InfoSphere Warehouse Data Mining während der Modellerstellung ein Feedback erhalten und dass SPSS Modeler die Modellerstellung abbrechen kann. Beachten Sie, dass SPSS Modeler durch diesen Konfigurationsschritt in die Lage versetzt wird, DB2-Daten zu lesen, die nicht an die Datenbank gebunden sind, indem Transaktionen gleichzeitig ausgeführt werden. Wenden Sie sich an Ihren Datenbankadministrator, wenn Sie sich über die Auswirkungen dieser Änderung im Unklaren sind.

Abbildung 5-1

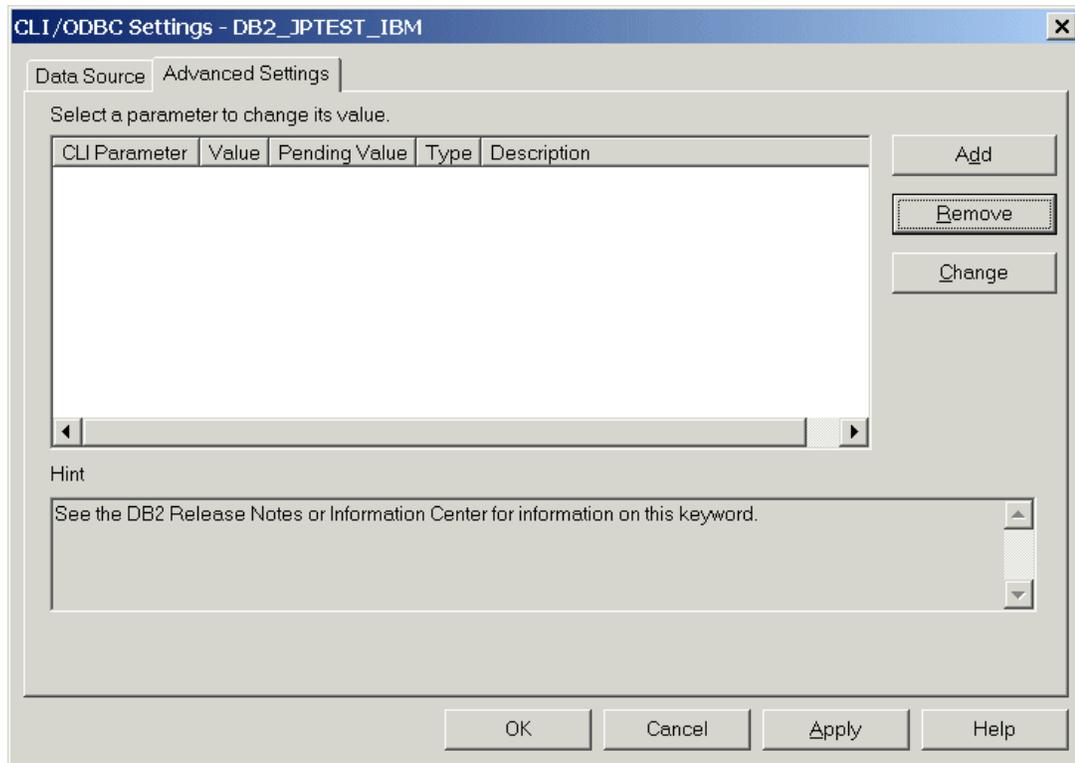
Dialogfeld "ODBC DB2 Wire Protocol Driver Setup," Registerkarte "Advanced"



**SPSS OEM 6.0 DB2 Wire Protocol-Treiber** Führen Sie für den Connect ODBC-Treiber die folgenden Schritte durch:

- ▶ Starten Sie den ODBC-Datenquellen-Administrator, wählen Sie die im vorangegangenen Abschnitt erstellte Datenquelle aus und klicken Sie auf die Schaltfläche Konfigurieren.
- ▶ Klicken Sie im Fenster zum Einrichten des ODBC DB2 Wire Protocol-Treibers auf Erweitert:
- ▶ Legen Sie als Standardisolationsebene 0-READ UNCOMMITTED fest und klicken Sie auf OK.

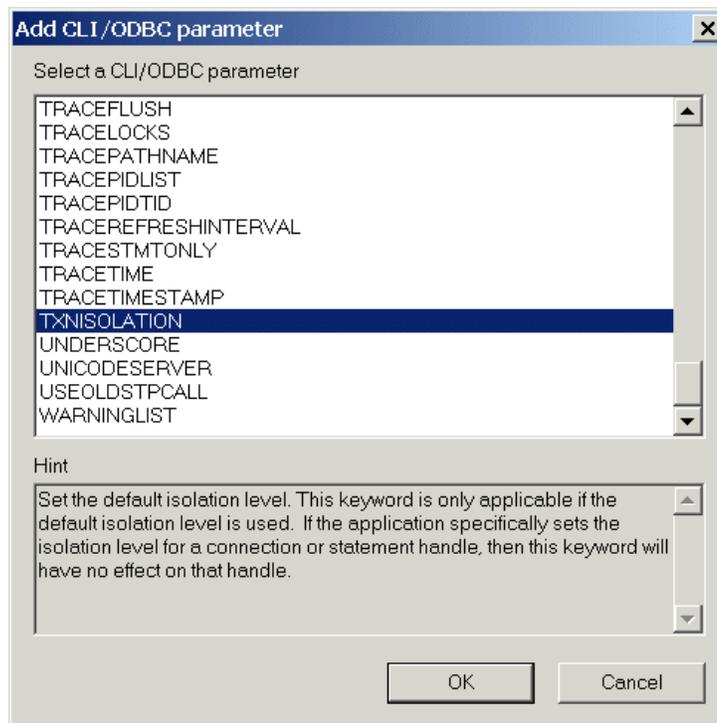
Abbildung 5-2  
Dialogfeld "CLI/ODBC Settings," Registerkarte "Advanced Settings"



**IBM DB2 ODBC Driver.** Führen Sie für den IBM DB2-Treiber die folgenden Schritte durch:

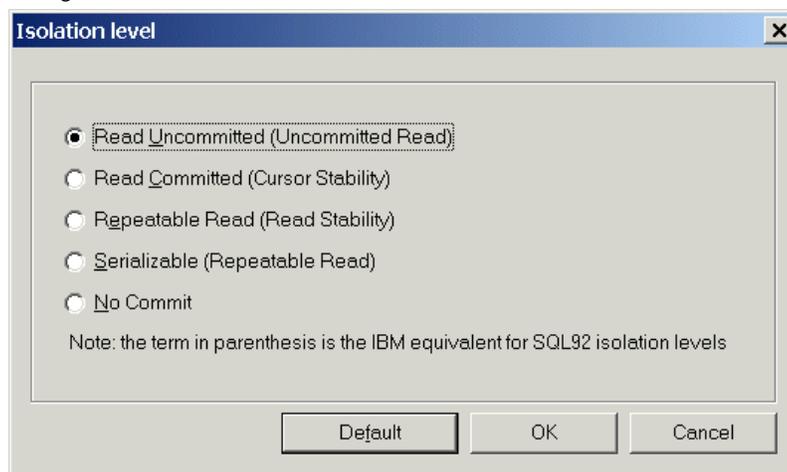
- ▶ Starten Sie den ODBC-Datenquellen-Administrator, wählen Sie die im vorangegangenen Abschnitt erstellte Datenquelle aus und klicken Sie auf die Schaltfläche Konfigurieren.
- ▶ Klicken Sie im Dialogfeld "CLI/ODBC Settings" auf die Registerkarte Advanced Settings (Erweiterte Einstellungen) und dann auf die Schaltfläche Add (Hinzufügen).

Abbildung 5-3  
Dialogfeld für CLI/ODBC-Parameter



- ▶ Wählen Sie im Dialogfeld “CLI/ODBC-Parameter” den Parameter TXNISOLATION und klicken Sie auf OK.

Abbildung 5-4  
Dialogfeld “Isolation Level”



- ▶ Wählen Sie im Dialogfeld “Isolation Level” (Isolationsgrad) den Parameter Read Uncommitted (Lesen nicht zugesichert) und klicken Sie auf OK.

- Klicken Sie im Dialogfeld mit den CLI/ODBC-Einstellungen auf OK, um die Konfiguration abzuschließen.

Beachten Sie, dass das von InfoSphere Warehouse Data Mining erstellte Feedback im folgenden Format übergeben wird:

<ITERATIONSNR> / <FORTSCHRITT> / <KERNELPHASE>

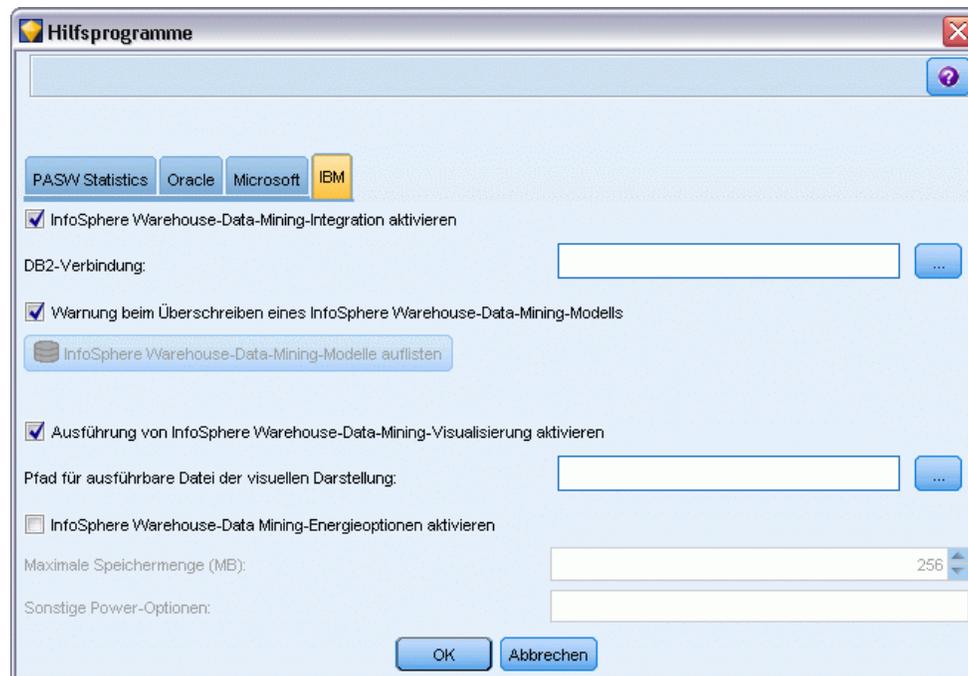
Dabei gilt:

- ITERATIONSNR ist die Anzahl der aktuell auf den Daten durchgeführten Durchläufe – beginnend mit 1.
- <FORTSCHRITT> zeigt den Fortschritt der aktuellen Iteration als Zahl zwischen 0,0 und 1,0 an.
- <KERNELPHASE> beschreibt die aktuelle Phase des Mining-Algorithmus.

### **Aktivieren der InfoSphere Warehouse Data Mining-Integration in IBM SPSS Modeler**

Damit SPSS Modeler DB2 mit InfoSphere Warehouse Data Mining benutzen kann, müssen Sie zuerst im Dialogfeld “Hilfsprogramme” einige Angaben vornehmen.

Abbildung 5-5  
Registerkarte “IBM” im Dialogfeld “Hilfsprogramme”



- Wählen Sie in den SPSS Modeler-Menüs folgende Befehlsfolge:  
Werkzeuge > Optionen > Hilfsprogramme
- Klicken Sie auf die Registerkarte IBM.

**Aktivieren der InfoSphere Warehouse Data Mining-Integration.** Aktiviert die Datenbank-Modellierungspalette, sofern diese nicht bereits angezeigt wird, und fügt Modellierungsknoten für ISW Data Mining-Integrationsalgorithmen einschließlich Entscheidungsbäumen, Assoziationen, Sequenz, Clusterbildung und Regression hinzu. (Nach der Aktivierung wird diese Palette zusammen mit den anderen am unteren Rand des SPSS Modeler-Fensters angezeigt.)

**DB2-Verbindung.** Legt die DB2 ODBC-Datenquelle fest, die bei der Bildung und dem Speichern der Modelle als Standard benutzt wird. Bei der konkreten Modellbildung und für die generierten Modellknoten kann diese Einstellung überschrieben werden. Klicken Sie auf die Schaltfläche mit den Auslassungspunkten (...), um die Datenquelle auszuwählen.

Die für die Modellbildung benutzte Datenbankverbindung kann, muss aber nicht mit der für den Datenzugriff benutzten übereinstimmen. Sie können beispielsweise einen Stream besitzen, der auf die Daten einer DB2-Datenbank zugreift, die Daten für die Bereinigung und sonstige Bearbeitungen in SPSS Modeler herunterlädt und dann zur Modellbildung in eine andere DB2-Datenbank lädt. Alternativ können sich die Originaldaten in einer Textdatei oder einer anderen Nicht-DB2-Quelle befinden. In diesem Fall müssen die Daten zur Modellbildung in DB2 geladen werden. In allen Fällen werden die Daten automatisch in eine temporäre Tabelle geladen, die bei Bedarf in der für die Modellbildung genutzten Datenbank angelegt wird.

**Warnen, wenn ein InfoSphere Warehouse Data Mining-Integrationsmodell überschrieben würde.** Wählen Sie diese Option, um sicherzustellen, dass in der Datenbank gespeicherte Modelle nicht von SPSS Modeler überschrieben werden, ohne dass eine Warnung ausgegeben wird.

**Aktivieren der InfoSphere Warehouse Data Mining-Modelle.** Mit dieser Option können Sie die in DB2 gespeicherten Modelle auflisten und löschen. [Für weitere Informationen siehe Thema Auflistung der Datenbankmodelle auf S. 117.](#)

**Start der Visualisierung von InfoSphere Warehouse Data Mining aktivieren.** Wenn Sie das Visualisierungsmodul installiert haben, müssen Sie es hier für die Verwendung in SPSS Modeler aktivieren.

**Pfad für ausführbare Datei der visuellen Darstellung** Die Position der ausführbaren Datei des Visualisierungsmoduls (falls installiert), z. B. *C:\Program Files\IBM\ISWarehouse\Im\IMVisualization\bin\imvisualizer.exe*.

**Zeitreihenvisualisierungs-Plugin-Verzeichnis** Die Position des Flash-Plugins für Zeitreihenvisualisierung (falls installiert), z. B. *C:\Program Files\IBM\ISWShared\plugins\com.ibm.datatools.datamining.imvisualization.flash\_2.2.1.v20091111\_0915*.

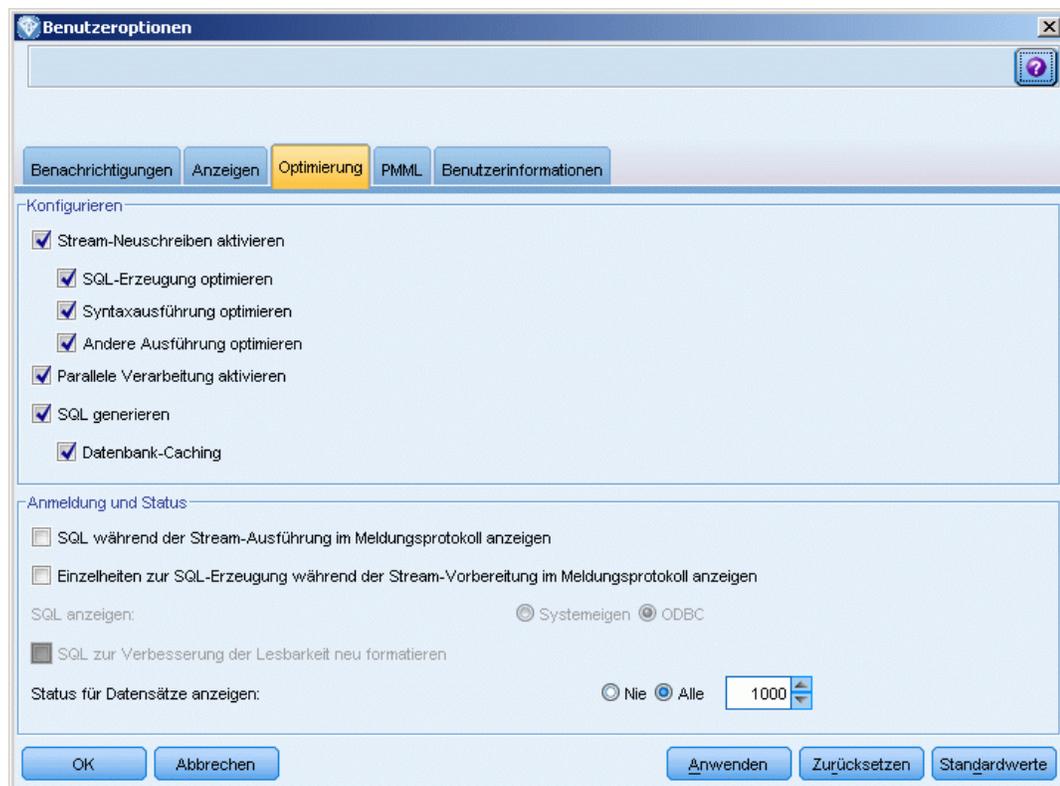
**Aktivieren der InfoSphere Warehouse Data Power-Optionen.** Sie können für den Algorithmus der datenbankinternen Modellbildung eine Obergrenze für die Arbeitsspeichernutzung einstellen und beliebige Befehlszeilenoptionen für bestimmte Modelle festlegen. Mit dem Speicherlimit können Sie den Speicherverbrauch steuern und einen Wert für die Power-Option `-buf` angeben. Weitere Power-Optionen können Sie hier in Befehlszeilenform angeben, die dann an InfoSphere Warehouse Data Mining übergeben werden. [Für weitere Informationen siehe Thema Power-Optionen auf S. 121.](#)

**InfoSphere Warehouse-Version überprüfen.** Überprüft die Version von InfoSphere Warehouse, die Sie verwenden, und meldet einen Fehler, wenn Sie versuchen eine Data-Mining-Funktion zu verwenden, die von Ihrer Version nicht unterstützt wird.

### **Aktivieren der SQL-Erzeugung und -Optimierung**

- ▶ Wählen Sie in den SPSS Modeler-Menüs folgende Befehlsfolge:  
Werkzeuge > Optionen > Benutzeroptionen

Abbildung 5-6  
Optimierungseinstellungen



- ▶ Klicken Sie dann auf die Registerkarte Optimierung.
- ▶ Überzeugen Sie sich, dass die Option SQL generieren aktiviert ist. Diese Einstellung ist für die Datenbank-Modellierung erforderlich.
- ▶ Wählen Sie SQL-Erzeugung optimieren und Andere Ausführung optimieren aus. (Diese Einstellungen sind nicht unbedingt erforderlich, werden aber zur Leistungsoptimierung empfohlen.)

Für weitere Informationen siehe Thema Festlegen von Optimierungsoptionen in Kapitel 12 in *IBM SPSS Modeler 14.2- Benutzerhandbuch*.

## Modellbildung mit InfoSphere Warehouse Data Mining

Für die Modellbildung von InfoSphere Warehouse Data Mining muss sich das Trainings-Daten-Set in einer Tabelle oder Ansicht innerhalb der DB2-Datenbank befinden. Wenn sich die Daten nicht in DB2 befinden oder wenn die Daten wegen einer Datenvorbereitung, die nicht in DB2 erfolgen kann, in IBM® SPSS® Modeler verarbeitet werden müssen, werden diese vor der Modellbildung automatisch in eine temporäre DB2-Tabelle geladen.

### Modell-Scoring und -Deployment

Das Modell-Scoring erfolgt immer innerhalb von DB2 und wird immer von InfoSphere Warehouse Data Mining durchgeführt. Wenn die Daten aus IBM® SPSS® Modeler stammen oder dort vorbereitet werden müssen, muss das Daten-Set gegebenenfalls in eine temporäre Tabelle geladen werden. Für Modelle vom Typ "Entscheidungsbaum", "Regression" und "Cluster-Bildung" in SPSS Modeler wird in der Regel nur eine Vorhersage mit der zugehörigen Wahrscheinlichkeit oder Konfidenz erstellt. Außerdem ist zum Anzeigen der Konfidenzen jedes möglichen Ergebnisses (ähnlich wie bei einer logistischen Regression) auf der Registerkarte "Einstellungen" des Modell-Nuggets eine Zeitoption für das Scoring (Kontrollkästchen Konfidenzen für alle Klassen einschließen) verfügbar. Bei Assoziations- und Sequenzmodellen in SPSS Modeler werden mehrere Werte ausgegeben. SPSS Modeler kann InfoSphere Warehouse Data Mining-Modelle mit IBM® SPSS® Modeler Solution Publisher aus zur Ausführung veröffentlichten Streams heraus scoren.

Folgende Felder werden durch Scoring-Modelle generiert:

Tabelle 5-1  
Modell-Scoring-Felder

Modelltyp	Score-Spalten	Bedeutung
Decision Trees (Entscheidungsbäume)	\$I-Feld	Beste Vorhersage für <i>Feld</i> .
	\$IC-Feld	Konfidenz der besten Vorhersage für <i>Feld</i> .
	\$IC-Wert1, ..., \$IC-WertN	(optional) Konfidenz für jeden von <i>N</i> möglichen Werten für <i>Feld</i> .
Regression	\$I-Feld	Beste Vorhersage für <i>Feld</i> .
	\$IC-Feld	Konfidenz der besten Vorhersage für <i>Feld</i> .
Clusterbildung	\$I-Modellname	Beste Cluster-Zuordnung für Eingabedatensatz.
	\$IC-Modellname	Konfidenz der besten Cluster-Zuordnung für Eingabedatensatz.

<b>Modelltyp</b>	<b>Score-Spalten</b>	<b>Bedeutung</b>
Assoziation	<i>\$I-Modellname</i>	ID der Übereinstimmungsregel.
	<i>\$IH-Modellname</i>	Kopfelement.
	<i>\$IHN-Modellname</i>	Name des Kopfelements.
	<i>\$IS-Modellname</i>	Unterstützungswert der Übereinstimmungsregel.
	<i>\$IC-Modellname</i>	Konfidenzwert der Übereinstimmungsregel.
	<i>\$IL-Modellname</i>	Lift-Wert der Übereinstimmungsregel.
	<i>\$IMB-Modellname</i>	Anzahl der übereinstimmenden Haupttextelemente oder Sets von Haupttextelementen (da alle Haupttextelemente bzw. Sets von Haupttextelementen mit dieser Zahl übereinstimmen müssen, ist sie gleich der Anzahl der Haupttextelemente bzw. Sets von Haupttextelementen).
Sequenz	<i>\$I-Modellname</i>	ID der Übereinstimmungsregel
	<i>\$IH-Modellname</i>	Set von Kopfelementen der Übereinstimmungsregel
	<i>\$IHN-Modellname</i>	Namen der Elemente im Kopfelemente-Set der Übereinstimmungsregel
	<i>\$IS-Modellname</i>	Unterstützungswert der Übereinstimmungsregel
	<i>\$IC-Modellname</i>	Konfidenzwert der Übereinstimmungsregel
	<i>\$IL-Modellname</i>	Lift-Wert der Übereinstimmungsregel
	<i>\$IMB-Modellname</i>	Anzahl der übereinstimmenden Haupttextelemente oder Sets von Haupttextelementen (da alle Haupttextelemente bzw. Sets von Haupttextelementen mit dieser Zahl übereinstimmen müssen, ist sie gleich der Anzahl der Haupttextelemente bzw. Sets von Haupttextelementen)
Naive Bayes	<i>\$I-Feld</i>	Beste Vorhersage für <i>Feld</i> .
	<i>\$IC-Feld</i>	Konfidenz der besten Vorhersage für <i>Feld</i> .
Logistische Regression	<i>\$I-Feld</i>	Beste Vorhersage für <i>Feld</i> .
	<i>\$IC-Feld</i>	Konfidenz der besten Vorhersage für <i>Feld</i> .

## **Verwalten von DB2-Modellen**

Bei der Erstellung eines InfoSphere Warehouse Data Mining-Modells mit IBM® SPSS® Modeler wird in SPSS Modeler ein Modell erzeugt und außerdem in der DB2-Datenbank ein Modell erzeugt oder ersetzt. Dieses SPSS Modeler-Modell stellt einen Bezug zum Inhalt eines in einem Datenbankserver gespeicherten Datenbankmodells her. SPSS Modeler kann eine Konsistenzprüfung durchführen, indem sowohl im SPSS Modeler-Modell als auch im DB2-Modell eine identische, generierte Modellschlüsselzeichenkette gespeichert wird.

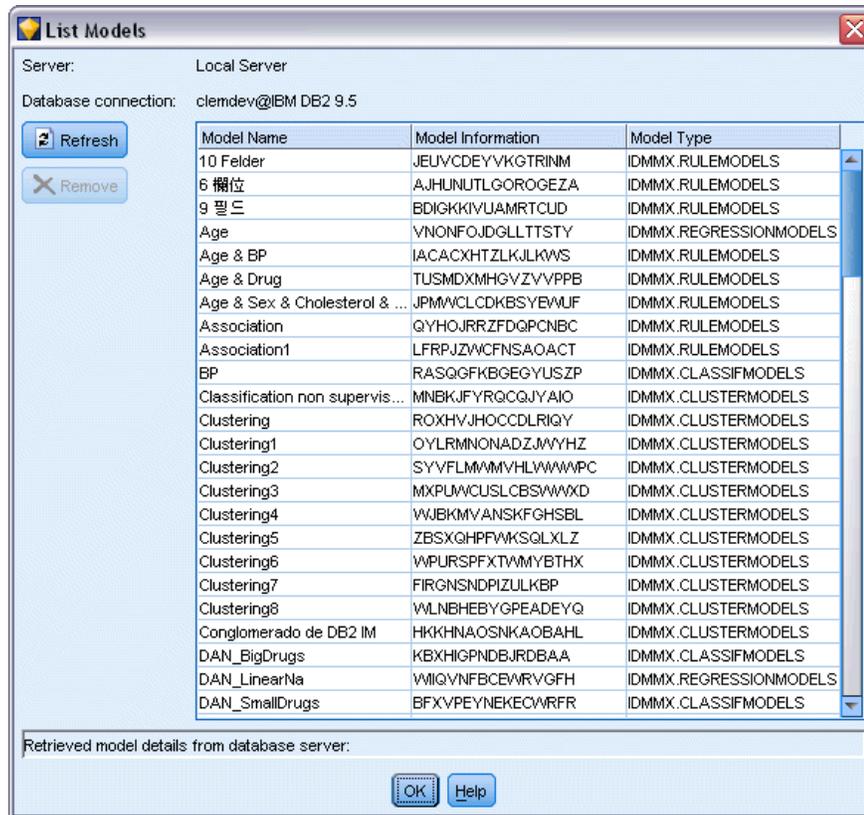
Die Schlüsselzeichenkette wird für jedes DB2-Modell im Dialogfeld "Listing Database Models" in der Spalte *Modellinformationen* angezeigt. Die Schlüsselzeichenkette eines SPSS Modeler-Modells wird auf der Registerkarte "Server" eines SPSS Modeler-Modells (wenn es sich in einem Stream befindet) als der Modellschlüssel ausgegeben.

Mit der Schaltfläche "Überprüfen" wird ermittelt, ob die Modellschlüssel des SPSS Modeler-Modells und des DB2-Modells übereinstimmen. Wenn in DB2 kein Modell mit demselben Namen gefunden wird oder wenn der Modellschlüssel nicht übereinstimmt, bedeutet dies, dass das DB2-Modell seit der Erstellung des SPSS Modeler-Modells gelöscht oder neu erstellt wurde. [Für weitere Informationen siehe Thema ISW-Modell-Nugget – Registerkarte "Server" auf S. 151.](#)

## **Auflistung der Datenbankmodelle**

IBM® SPSS® Modeler bietet ein Dialogfeld, in dem die in InfoSphere Warehouse Data Mining gespeicherten Modelle aufgelistet und gelöscht werden können.

Abbildung 5-7  
DB2 List Models (Dialogfeld)



Der Zugriff auf dieses Dialogfeld erfolgt über das Dialogfeld “IBM Helper Applications” und über die Dialogfelder zur Modellbildung, zum Suchen und zum Anwenden für die mit InfoSphere Warehouse Data Mining verbundenen Knoten. Zu jedem Modell werden folgende Informationen angezeigt:

- Modellname (Name des Modells – zum Sortieren der Liste verwendet)
- Modellinformationen (Modellschlüsselinformation aus einem bei der Erstellung des Modells von SPSS Modeler erzeugten Zufallsschlüssel)
- Modelltyp (die DB2-Tabelle, in der InfoSphere Warehouse Data Mining das Modell gespeichert hat)

### **Durchsuchen von Modellen**

Das Visualisierungs-Tool stellt die einzige Methode zum Durchsuchen von InfoSphere Warehouse Data Mining-Modellen dar. Das Tool kann optional mit InfoSphere Warehouse Data Mining installiert werden. [Für weitere Informationen siehe Thema Aktivieren der Integration mit InfoSphere Warehouse auf S. 107.](#)

- Klicken Sie auf Ansicht, um das Visualizer-Tool zu starten. Was das Tool anzeigt, hängt vom generierten Knotentyp ab. Beispielsweise gibt das Visualizer-Tool eine Ansicht der vorhergesagten Klassen aus, wenn es von einem ISW Entscheidungsbaum-Modell-Nugget gestartet wird.
- Klicken Sie auf Testergebnisse (nur Entscheidungsbäume und Sequenz), um das Visualizer-Tool zu starten und die Gesamtqualität des generierten Modells anzuzeigen.

### **Exportieren von Modellen und Generieren von Knoten**

Sie können für InfoSphere Warehouse Data Mining-Modelle einen PMML-Import und -Export durchführen. Die exportierte PMML-Datei enthält das von InfoSphere Warehouse Data Mining generierte Original-PMML. Die Exportfunktion bringt das Modell wieder in das PMML-Format.

Sie können eine Modellübersicht und -struktur als Text- oder HTML-Datei exportieren. Sie können die benötigten Filter-, Auswahl und Ableitungsknoten generieren. Weitere Informationen finden Sie im *IBM® SPSS® Modeler-Benutzerhandbuch* unter "Exportieren von Modellen".

### **Knoteneinstellungen, die für alle Algorithmen gelten**

Folgende Einstellungen haben viele der InfoSphere Warehouse Data Mining-Algorithmen gemeinsam:

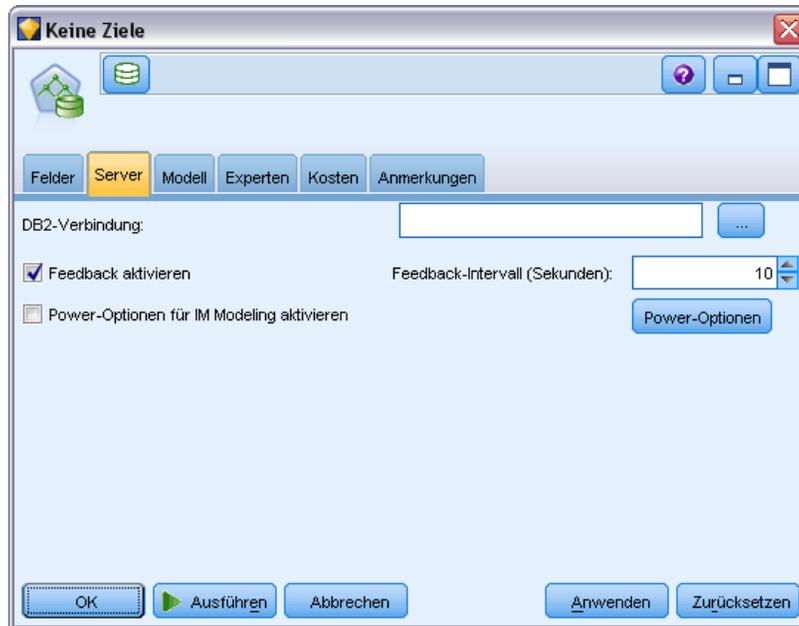
**Ziel und Prädiktoren.** Ein Ziel und Prädiktoren legen Sie entweder über den Typknoten fest oder manuell auf der Registerkarte "Felder" des Modellierungsknotens. Letzteres ist die Standardvorgehensweise in IBM® SPSS® Modeler.

**ODBC-Datenquellen.** Mit dieser Einstellung kann der Benutzer die Standard-ODBC-Datenquelle für das aktuelle Modell überschreiben. (Der Standard wird im Dialogfeld "Hilfsprogramme" festgelegt.) [Für weitere Informationen siehe Thema Aktivieren der Integration mit InfoSphere Warehouse auf S. 107.](#)

### **Optionen der Registerkarte "ISW-Server"**

Sie können festlegen, welche DB2-Verbindung zum Hochladen der für die Modellbildung verwendeten Daten benutzt wird. Gegebenenfalls können Sie auf der Registerkarte "Server" für jeden Modellierungsknoten eine Verbindung auswählen, mit der die im Dialogfeld "Hilfsprogramme" angegebene Standard-DB2-Verbindung überschrieben wird. [Für weitere Informationen siehe Thema Aktivieren der Integration mit InfoSphere Warehouse auf S. 107.](#)

Abbildung 5-8  
Registerkarte "ISW-Server"



Die für die Modellierung benutzte Verbindung kann mit der im Quellenknoten für einen Stream benutzen Verbindung identisch sein. Sie können beispielsweise einen Stream besitzen, der auf die Daten einer DB2-Datenbank zugreift, die Daten für die Bereinigung und sonstige Bearbeitungen in IBM® SPSS® Modeler herunterlädt und dann zur Modellbildung in eine andere DB2-Datenbank lädt.

Der Name der ODBC-Datenquelle wird in jeden SPSS Modeler-Stream eingebettet. Wenn ein auf einem Host erzeugter Stream auf einem anderen Host ausgeführt wird, muss der Name der Datenquelle auf beiden Hosts identisch sein. Alternativ kann für jeden Quellen- oder Modellierungsknoten auf der Registerkarte "Server" eine andere Datenquelle ausgewählt werden.

Mit den folgenden Optionen erhalten Sie beim Erstellen eines Modells Feedback:

- **Feedback aktivieren.** Wählen Sie diese Option, damit Sie während der Modellbildung Feedback erhalten (standardmäßig ausgeschaltet).
- **Feedback-Intervall (Sekunden).** Legen Sie fest, wie oft SPSS Modeler Feedback über den Fortschritt der Modellbildung abrufen.

**Aktivieren der InfoSphere Warehouse Data Power-Optionen.** Wählen Sie diese Option, um die Schaltfläche Power-Optionen zu aktivieren, die Ihnen ermöglicht, eine Reihe erweiterter Optionen wie ein Speicherlimit und benutzerdefinierte SQL anzugeben. [Für weitere Informationen siehe Thema Power-Optionen auf S. 121.](#)

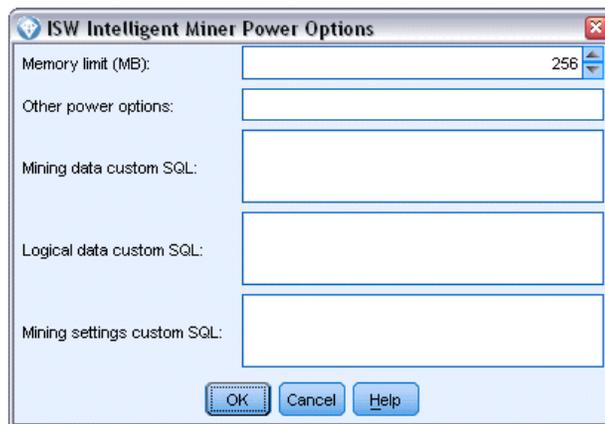
Die Registerkarte "Server" eines generierten Knotens enthält eine Option für die Durchführung einer Konsistenzprüfung, für die sowohl im SPSS Modeler-Modell als auch im DB2-Modell eine identisch erzeugte Modellschlüsselzeichenkette gespeichert wird. [Für weitere Informationen siehe Thema ISW-Modell-Nugget – Registerkarte "Server" auf S. 151.](#)

## Power-Optionen

Auf der Registerkarte "Server" für alle vier Algorithmen befindet sich ein Kontrollkästchen, mit dem die Power-Optionen für ISW Modeling aktiviert werden. Wenn Sie auf die Schaltfläche Power-Optionen klicken, wird der Bildschirm "Power-Optionen für ISW" angezeigt, der folgende Optionen bietet:

- Maximale Speichermenge (MB).
- Sonstige Power-Optionen.
- Benutzerdefinierte SQL für Mining-Daten.
- Benutzerdefinierte SQL für logische Daten.
- Benutzerdefinierte SQL für Mining-Einstellungen.

Abbildung 5-9  
Einstellungen der Power-Optionen für ISW



**Maximale Speichermenge (MB).** Beschränkt den Speicherverbrauch eines Modellbildungsalgorithmus. Beachten Sie, dass die Standard-Power-Option eine Obergrenze der Anzahl diskreter Werte in kategorialen Daten festlegt.

**Sonstige Power-Optionen.** Hier können beliebige Power-Optionen in Befehlszeilenform angegeben werden, die für bestimmte Modelle oder Lösungen gelten. Die Angaben fallen je nach Implementierung oder Lösung unterschiedlich aus. Um eine Modellbildungsaufgabe zu definieren, können Sie die von IBM® SPSS® Modeler generierte SQL manuell erweitern.

**Benutzerdefinierte SQL für Mining-Daten.** Sie können Methodenaufrufe hinzufügen, um das Objekt `DM_MiningData` anzupassen. Mit der folgenden SQL wird beispielsweise zu den bei der Modellbildung benutzten Daten ein Filter hinzugefügt, der auf einem Feld namens *Partition* basiert:

```
..DM_setWhereClause('Partition' = 1')
```

**Benutzerdefinierte SQL für logische Daten.** Sie können Methodenaufrufe hinzufügen, um das Objekt `DM_LogicalDataSpec` anzupassen. Die folgende SQL entfernt beispielsweise ein Feld aus dem für die Modellbildung benutzten Feld-Set:

```
..DM_remDataSpecFld('field6')
```

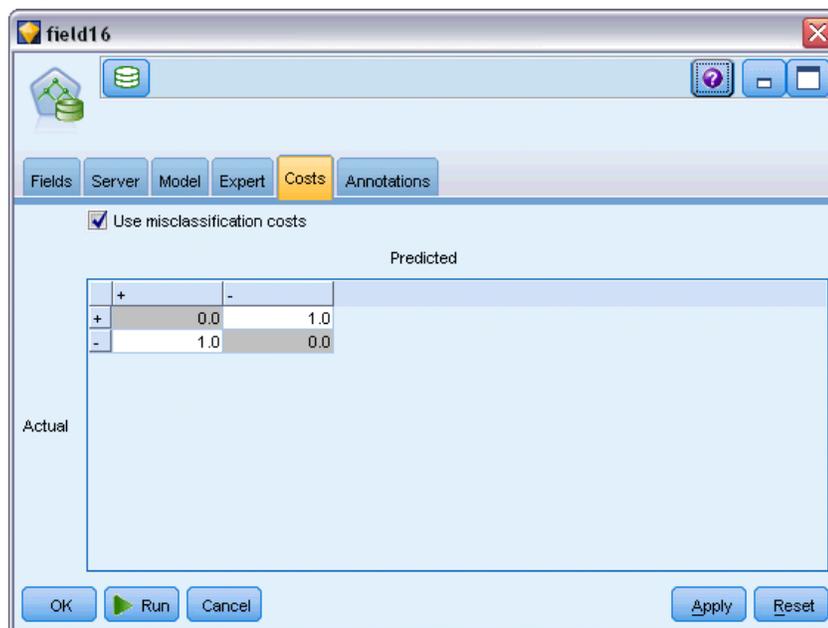
**Benutzerdefinierte SQL für Mining-Einstellungen.** Sie können Methodenaufrufe hinzufügen, um das Objekt `DM_ClasSettings/DM_RuleSettings/DM_ClusSettings/DM_RegSettings` anzupassen. Die Eingabe der folgenden SQL weist InfoSphere Warehouse Data Mining beispielsweise an, das Feld `Partition` zu aktivieren (dies bedeutet, dass es immer im resultierenden Modell enthalten ist):

```
..DM_setFldUsageType('Partition',1)
```

### Kostenoptionen für ISW

Auf der Registerkarte “Kosten” können Sie die Fehlklassifizierungskosten anpassen, mit denen Sie die relative Wichtigkeit verschiedener Arten von Vorhersagefehlern festlegen können.

Abbildung 5-10  
Registerkarte “Kosten” für ISW



In einigen Zusammenhängen sind bestimmte Fehler kostspieliger als andere. Zum Beispiel kann es kostspieliger sein, einen Kreditantragsteller mit hohem Risiko als niedriges Risiko zu klassifizieren (eine Art von Fehler) als einen Kreditantragsteller mit niedrigem Risiko als hohes Risiko (eine andere Art von Fehler). Anhand von Fehlklassifizierungskosten können Sie die relative Bedeutung verschiedener Arten von Vorhersagefehlern angeben.

Fehlklassifizierungskosten sind im Grunde Gewichtungen, die auf bestimmte Ergebnisse angewendet werden. Diese Gewichtungen werden in das Modell mit einbezogen und können die Vorhersage (als Schutz gegen kostspielige Fehler) ändern.

Mit Ausnahme von C5.0-Modellen werden Fehlklassifizierungskosten beim Scoring von Modellen nicht angewendet und bei der Rangbildung bzw. beim Vergleich von Modellen mithilfe eines Knotens vom Typ “Automatischer Klassifizierer”, eines Evaluationsdiagramms oder eines Analyseknosens nicht berücksichtigt. Ein Modell, das Kosten beinhaltet, führt nicht unbedingt zu weniger Fehlern als ein Modell, das keine Kosten beinhaltet, und es weist auch nicht

unbedingt eine höhere Gesamtgenauigkeit auf, aber es ist möglicherweise in praktischer Hinsicht leistungsfähiger, da es eine integrierte Verzerrung zugunsten von *weniger teuren* Fehlern aufweist.

Die Kostenmatrix zeigt die Kosten für jede mögliche Kombination aus prognostizierter Kategorie und tatsächlicher Kategorie. Standardmäßig werden alle Fehlklassifizierungskosten auf 1,0 gesetzt. Um benutzerdefinierte Kostenwerte einzugeben, wählen Sie Fehlklassifizierungskosten verwenden und geben Ihre benutzerdefinierten Werte in die Kostenmatrix ein.

Um die Fehlklassifizierungskosten zu ändern, wählen Sie die Zelle aus, die der gewünschten Kombination aus vorhergesagten und tatsächlichen Werten entspricht, löschen den Inhalt der Zelle und geben die gewünschten Kosten für die Zelle ein. Die Kosten sind nicht automatisch symmetrisch. Wenn Sie z. B. die Kosten einer Fehlklassifizierung von *A* als *B* auf 2,0 setzen, weisen die Kosten für eine Fehlklassifizierung von *B* als *A* weiterhin den Standardwert 1,0 auf, es sei denn, Sie ändern diesen Wert ausdrücklich.

## **ISW-Entscheidungsbaum**

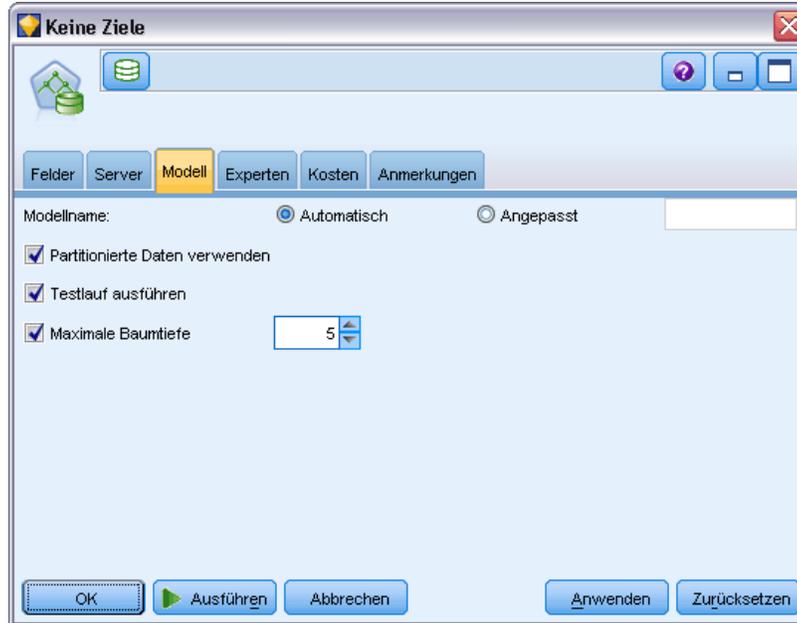
Mithilfe von Entscheidungsbaummodellen werden Klassifizierungssysteme entwickelt, die zukünftige Beobachtungen auf der Grundlage eines Satzes von Entscheidungsregeln vorhersagen oder klassifizieren. Wenn die Daten in Klassen aufgeteilt sind, die Sie interessieren (z. B. Darlehen mit hohem Risiko im Gegensatz zu Darlehen mit niedrigem Risiko, Abonnenten gegenüber Personen ohne Abonnement, Wähler im Gegensatz zu Nichtwählern oder Bakterienarten), können Sie mit diesen Daten Regeln erstellen, die Sie zur Klassifizierung alter oder neuer Fälle mit maximaler Genauigkeit verwenden können. So können Sie z. B. einen Baum erstellen, der das Kreditrisiko oder die Kaufabsicht basierend auf Alter und anderen Faktoren klassifiziert.

Der ISW-Entscheidungsbaum-Algorithmus erzeugt Klassifizierungsbäume aus kategorialen Eingabedaten. Der resultierende Entscheidungsbaum ist binär. Für die Bildung des Modells können eine Vielzahl von Einstellungen, einschließlich der Fehlklassifizierungskosten, vorgenommen werden.

Das ISW-Visualisierungs-Tool stellt die einzige Methode zum Durchsuchen von InfoSphere Warehouse Data Mining-Modellen dar.

## Optionen für ISW-Entscheidungsbaummodelle

Abbildung 5-11  
ISW-Entscheidungsbaumknoten – Registerkarte "Modell"



**Modellname.** Sie können den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen (oder dem Modelltyp in Fällen, in denen kein solches Feld angegeben ist) generieren oder einen benutzerdefinierten Namen eingeben.

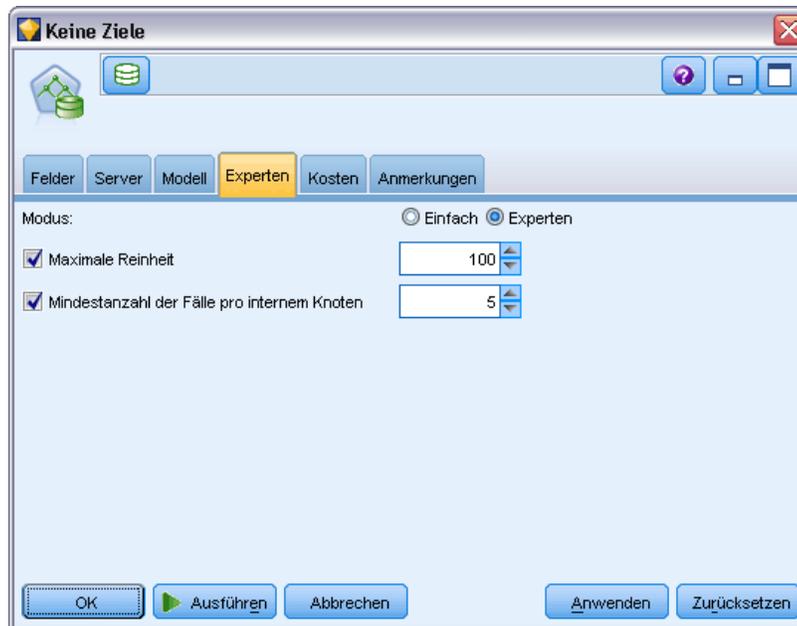
**Partitionierte Daten verwenden.** Wenn Sie ein Partitionsfeld definieren, wählen Sie Partitionierte Daten verwenden aus.

**Testlauf ausführen.** Sie können auswählen, dass ein Testlauf ausgeführt werden soll. Dann wird nach der Modellbildung in der Trainingspartition ein InfoSphere Warehouse Data Mining-Testlauf ausgeführt. Dabei wird ein Lauf über die Testpartition durchgeführt und es werden Modellqualitätsinformationen, Lift Charts etc. erzeugt.

**Maximale Baumtiefe.** Sie können die maximale Strukturtiefe festlegen. Dies schränkt die Tiefe des Baums auf die angegebene Anzahl Ebenen ein. Wenn diese Option nicht ausgewählt ist, wird keine Obergrenze vorgegeben. Um die Überlagerung komplexer Modelle zu vermeiden, ist selten ein über 5 liegender Wert empfehlenswert.

## Expertenoptionen für ISW-Entscheidungsäume

Abbildung 5-12  
ISW-Entscheidungsbaumknoten – Registerkarte "Experten"



**Maximale Reinheit.** Diese Option legt die maximale Reinheit für interne Knoten fest. Wenn das Aufteilen eines Knotens dazu führt, dass eines der untergeordneten Elemente den angegebenen Reinheitswert überschreitet (Bsp.: Mehr als 90 % der Fälle fallen unter eine festgelegte Kategorie), dann wird der Knoten nicht geteilt.

**Mindestanzahl der Fälle pro internen Knoten.** Wenn das Aufteilen eines Knotens dazu führt, dass ein Knoten mit weniger Fällen entsteht, als dies der Minimalwert vorgibt, dann wird der Knoten nicht aufgeteilt.

## ISW-Assoziation

Sie können den ISW Assoziationsknoten verwenden, um Assoziationsregeln zwischen Elementen zu ermitteln, die in einem Set von Gruppen vorhanden sind. Assoziationsregeln ordnen eine bestimmte Schlussfolgerung (beispielsweise den Kauf eines bestimmten Produkts) einer Menge von Bedingungen (dem Kauf mehrerer anderer Produkte) zu.

Sie können Assoziationsregeln nach Wunsch im Modell ein- oder ausschließen, indem Sie **Beschränkungen** festlegen. Wenn Sie ein bestimmtes Eingabefeld einschließen, werden Assoziationsregeln in das Modell übernommen, die mindestens eines der angegebenen Elemente enthalten. Wenn Sie ein Eingabefeld ausschließen, werden Assoziationsregeln, die eines der angegebenen Elemente enthalten, in den Ergebnissen nicht berücksichtigt.

Die Assoziations- und Sequenzalgorithmen von ISW können **Taxonomien** verwenden. Taxonomien bilden einzelne Werte auf Konzepte einer höheren Ebene ab. Beispiel: Kugelschreiber und Bleistift können auf eine gleich bleibende Kategorie abgebildet werden.

Assoziationsregeln haben ein einziges Sukzedens (die Schlussfolgerung) und mehrere Antezedenzen (das Set der Bedingungen). Ein Beispiel lautet folgendermaßen:

[Brot, Marmelade] • [Butter]

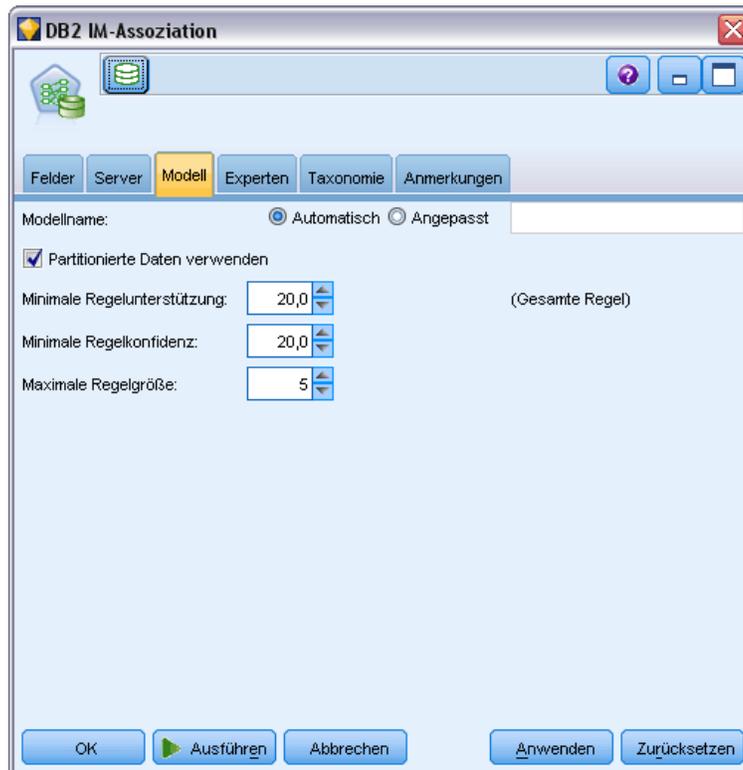
[Brot, Marmelade]  
• [Margarine]

Hier sind Bread und Jam die Antezedenzen (auch **Regeltext** genannt) und Butter oder Margarine sind Beispiele für ein Sukzedens (auch **Kopf der Regel** genannt). Die erste Regel gibt an, dass eine Person, die Brot und Marmelade gekauft hat, gleichzeitig auch Butter gekauft hat. Die zweite Regel identifiziert einen Kunden, der beim Einkauf der gleichen Kombination (Brot und Marmelade) beim selben Besuch des Geschäfts auch Margarine gekauft hat.

Das Visualisierungs-Tool stellt die einzige Methode zum Durchsuchen von InfoSphere Warehouse Data Mining-Modellen dar.

## Optionen für ISW-Assoziationsmodelle

Abbildung 5-13  
ISW-Assoziationsknoten – Registerkarte "Modell"



**Modellname.** Sie können den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen (oder dem Modelltyp in Fällen, in denen kein solches Feld angegeben ist) generieren oder einen benutzerdefinierten Namen eingeben.

**Partitionierte Daten verwenden.** Wenn ein Partitionsfeld definiert ist, gewährleistet diese Option, dass nur Daten aus der Trainingspartition für die Modellerstellung verwendet werden. [Für weitere Informationen siehe Thema Partitionsknoten in Kapitel 4 in IBM SPSS Modeler 14.2- Quellen-, Prozess- und Ausgabeknoten.](#)

**Minimale Regelunterstützung (%).** Minimales Unterstützungsniveau für Assoziations- oder Sequenzregeln. Nur Regeln, die mindestens dieses Unterstützungsniveau erreichen, werden in das Modell eingeschlossen. Der Wert wird als  $A/B*100$  berechnet, wobei A die Anzahl der Gruppen ist, die alle in der Regel vorkommenden Elemente enthalten, und B ist die Gesamtzahl aller berücksichtigten Gruppen. Wenn Sie weitere gemeinsame Assoziationen oder Sequenzen wünschen, erhöhen Sie diese Einstellung.

**Minimale Regelkonfidenz (%).** Minimales Konfidenzniveau für Assoziations- oder Sequenzregeln. Nur Regeln, die mindestens dieses Konfidenzniveau erreichen, werden in das Modell eingeschlossen. Der Wert wird als  $m/n*100$  berechnet. Dabei ist  $m$  die Anzahl der Gruppen, die den per Join verbundenen Kopf der Regel (Sukzedens) und Regeltext (Antezedens) enthalten, und  $n$  ist die Anzahl der Gruppen, die den Regeltext enthalten. Wenn Sie zu viele oder uninteressante Assoziationen oder Sequenzen erhalten, sollten Sie diese Einstellung erhöhen. Wenn Sie zu wenige Assoziationen oder Sequenzen erhalten, sollten Sie diese Einstellung reduzieren.

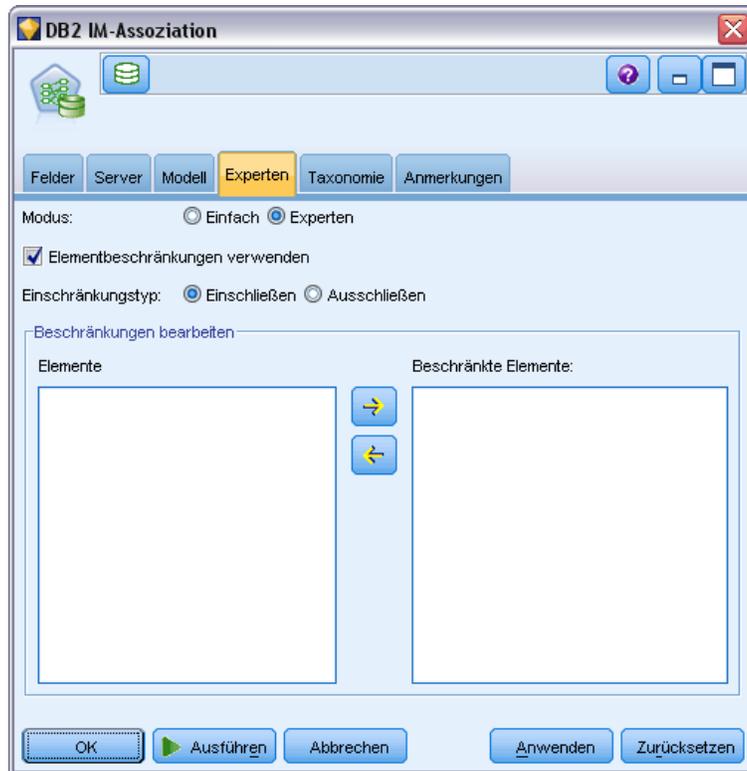
**Maximale Regelgröße.** Maximal zulässige Elemente in einer Regel, einschließlich des Sukzedens-Elements. Wenn die gewünschten Assoziationen oder Sequenzen relativ kurz sind, können Sie diese Einstellung reduzieren, um die Erstellung des Sets zu beschleunigen.

*Anmerkung:* Nur Knoten mit Transaktionseingabeformat werden gescort; Wahrheitstabellenformate (Tabellendaten) werden nicht verfeinert.

### **Expertenoptionen für ISW-Assoziationen**

Auf der Registerkarte "Experte" des Assoziationsknotens können Sie die Assoziationsregeln angeben, die in den Ergebnissen ein- oder ausgeschlossen werden sollen. Wenn Sie bestimmte Elemente einschließen, werden die Regeln in das Modell übernommen, die mindestens eines der angegebenen Elemente enthalten. Wenn Sie angegebene Elemente ausschließen, werden die Regeln, die eines der angegebenen Elemente enthalten, in den Ergebnissen nicht berücksichtigt.

Abbildung 5-14  
ISW-Assoziationsknoten – Registerkarte "Experten"



Wenn Elementbeschränkungen verwenden ausgewählt ist, werden alle Elemente, die Sie in die Beschränkungsliste eingefügt haben, abhängig von der Einstellung für den Einschränkungstyp, in den Ergebnissen ein- oder ausgeschlossen.

**Einschränkungstyp.** Wählen Sie, ob in den Ergebnissen diejenigen Assoziationsregeln ein- oder ausgeschlossen werden sollen, die die angegebenen Elemente enthalten.

**Beschränkungen bearbeiten.** Um ein Element zur Liste beschränkter Elemente hinzuzufügen, wählen Sie es in der Liste "Elemente" aus und klicken auf die Schaltfläche mit dem Rechtspfeil.

## ***ISW-Taxonomieoptionen***

Die Assoziations- und Sequenzalgorithmen von ISW können **Taxonomien** verwenden. Taxonomien bilden einzelne Werte auf Konzepte einer höheren Ebene ab. Beispiel: Kugelschreiber und Bleistift können auf eine gleich bleibende Kategorie abgebildet werden.

Auf der Registerkarte "Taxonomie" können Sie Kategoriezuordnungen definieren, mit denen Taxonomien innerhalb der Daten ausgedrückt werden. Beispiel: Eine Taxonomie erzeugt zwei Kategorien (Staple und Luxury) und ordnet Basiselemente dann einer dieser Kategorien zu. wine

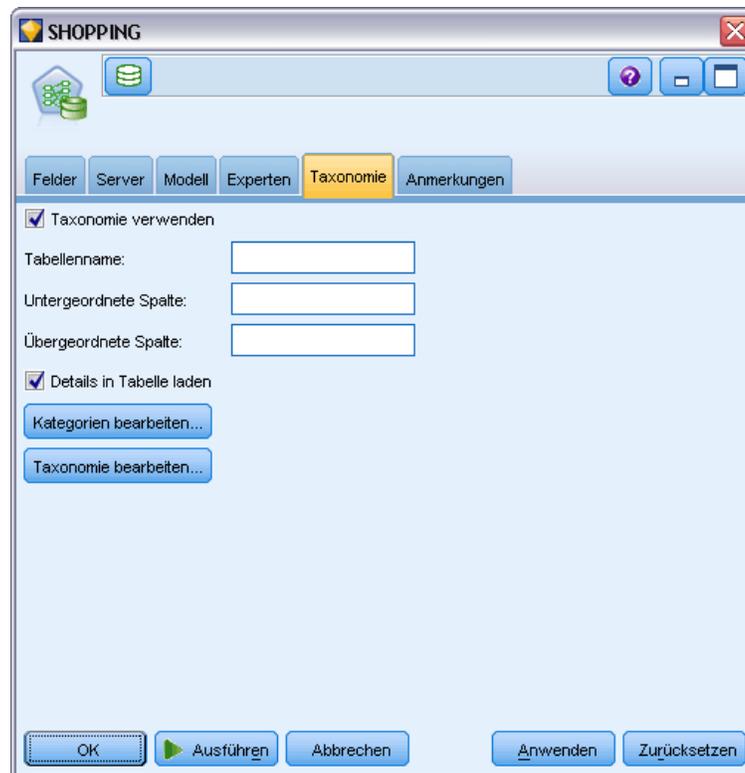
wird beispielsweise Luxury und bread wird Staple zugeordnet. Die Taxonomie verfügt über eine Parent-Child-Struktur wie folgt:

Kindobjekt	Elternobjekt
Wein	Luxus
Brot	Grundnahrungsmittel

Mit dieser Taxonomie können Sie ein Assoziations- oder Sequenzmodell bilden, das Regeln enthält, die sich sowohl auf die Kategorien als auch auf die Basiselemente beziehen.

*Hinweis:* Um die Optionen auf dieser Registerkarte zu aktivieren, müssen die Quelldaten in Transaktionsformat sein und Sie müssen Transaktionsformat verwenden in der Registerkarte Felder und anschließend Taxonomie verwenden in dieser Registerkarte auswählen. [Für weitere Informationen siehe Thema Tabellendaten im Vergleich zu Transaktionsdaten in Kapitel 12 in IBM SPSS Modeler 14.2-Modellierungsknoten.](#)

Abbildung 5-15  
ISW-Assoziationsknoten, Registerkarte "Taxonomie"



**Tabellenname.** Diese Option bestimmt den Namen der DB2-Tabelle, in der Taxonomiedetails gespeichert werden.

**Untergeordnete Spalte.** Diese Option bestimmt den Namen der untergeordneten Spalte in der Taxonomietabelle. Die untergeordnete Spalte enthält die Element- oder Kategorienamen.

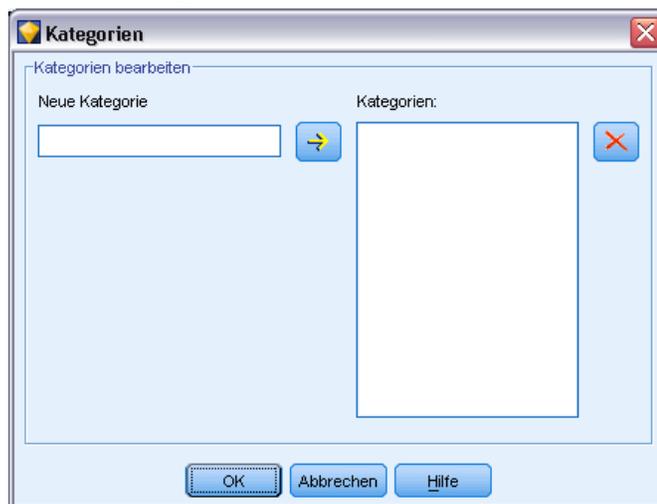
**Übergeordnete Spalte.** Diese Option bestimmt den Namen der übergeordneten Spalte in der Taxonomietabelle. Die übergeordnete Spalte enthält die Kategorienamen.

**Details in Tabelle laden.** Wählen Sie diese Option aus, wenn eine in IBM® SPSS® Modeler gespeicherte Taxonomieinformation zum Zeitpunkt der Modellbildung in die Taxonomietabelle geladen werden soll. Die Taxonomietabelle wird verworfen, wenn sie bereits vorhanden ist. Taxonomieinformationen werden mit dem Modellerstellungsknoten gespeichert; eine Bearbeitung erfolgt über die Schaltflächen “Kategorien bearbeiten” und “Taxonomie bearbeiten”.

### **Kategorieneditor**

Im Dialogfeld “Kategorien bearbeiten” können Sie Kategorien hinzufügen und aus einer sortierten Liste löschen.

Abbildung 5-16  
Taxonomie-Kategorieneditor



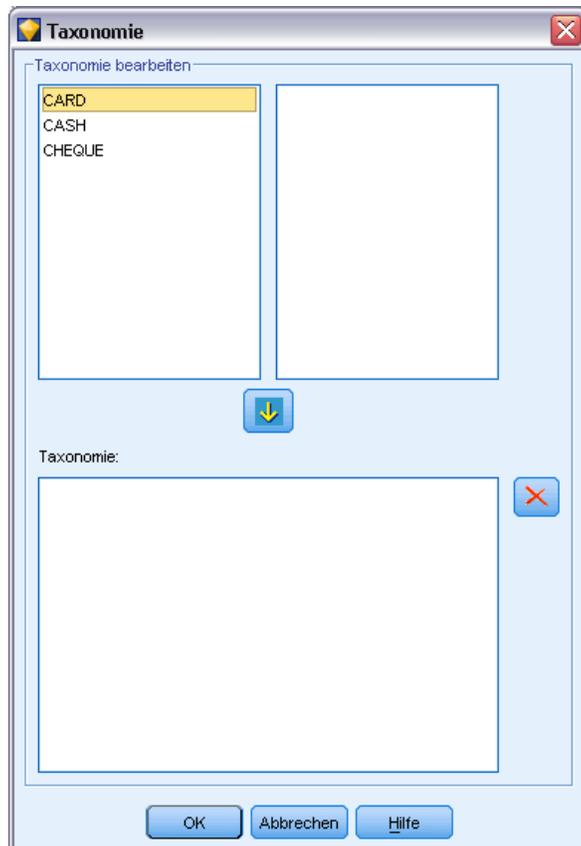
Sie fügen eine Kategorie hinzu, indem Sie ihren Namen in das Feld Neue Kategorie eingeben und auf die Pfeilschaltfläche klicken, um sie in die Liste Kategorien zu verschieben.

Eine Kategorie entfernen Sie, indem Sie sie in der Liste Kategorien auswählen und auf die benachbarte Schaltfläche “Löschen” klicken.

### **Taxonomieeditor**

Im Dialogfeld “Taxonomie bearbeiten” wird das Set der in den Daten definierten Basiselemente und das Set der Kategorien angegeben, die zu einer Taxonomie kombiniert werden. Um Einträge zur Taxonomie hinzuzufügen, wählen Sie aus der links angezeigten Liste beliebig viele Elemente oder Kategorien und aus der rechts angezeigten Liste eine oder mehrere Kategorien aus. Klicken Sie anschließend auf die Pfeilschaltfläche. Denken Sie daran, dass Kategorien nicht zu Taxonomien hinzugefügt werden, wenn dies zu einem Konflikt führt (Beispiel: Die Angabe `cat1` -> `cat2` und das Gegenteil `cat2` -> `cat1`).

Abbildung 5-17  
Taxonomieeditor



## ISW-Sequenz

Der Sequenzknoten erkennt Muster in sequenziellen oder zeitorientierten Daten, und zwar im Format Brot -> Käse. Die Elemente einer Sequenz sind **Element-Sets**, die eine einzelne Transaktion ausmachen. Beispiel: Wenn eine Person in den Supermarkt geht und Brot und Milch kauft und dann ein paar Tage später zurückkehrt und Käse kauft, kann das Kaufverhalten dieser Person als zwei Element-Sets dargestellt werden. Der erste Element-Set enthält Brot und Milch, der zweite Käse. Eine **Sequenz** ist eine Liste mit Element-Sets, die in einer vorhersagbaren Reihenfolge auftreten. Der Sequenzknoten erkennt häufige Sequenzen und erstellt einen generierten Modellknoten, der für Vorhersagen verwendet werden kann.

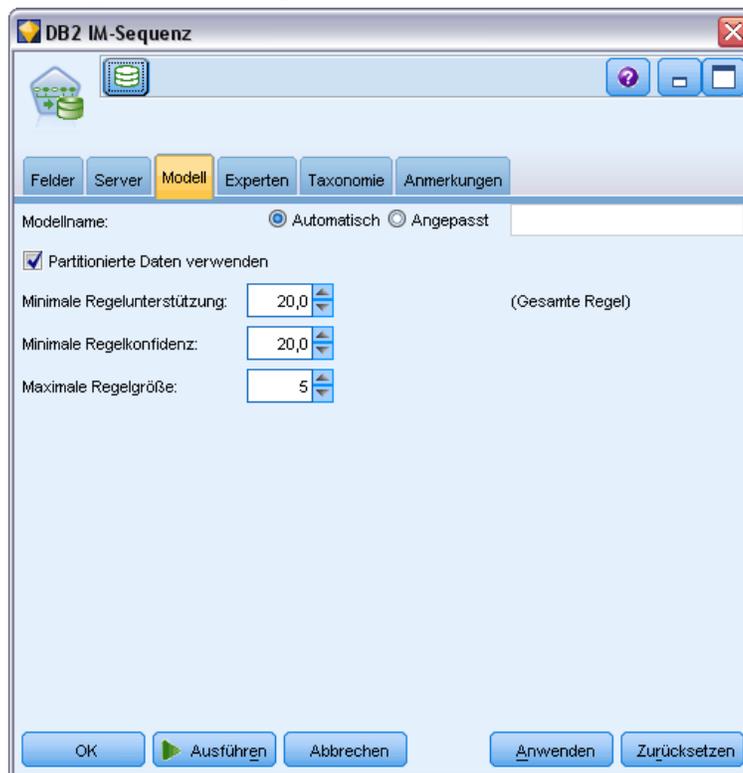
Sie können die Mining-Funktion für Sequenzregeln in verschiedenen Geschäftsbereichen verwenden. Im Einzelhandel beispielsweise können Sie typische Serien von Einkäufen ermitteln. Diese Serien zeigen Ihnen die verschiedenen Kombinationen von Kunden, Produkten und Einkaufszeitpunkt. Mit diesen Informationen können Sie potenzielle Kunden für ein bestimmtes Produkt identifizieren, die das Produkt noch nicht erworben haben. Des Weiteren können Sie den potenziellen Kunden rechtzeitig Produkte anbieten.

Eine Sequenz ist ein geordnetes Set von Element-Sets. Sequenzen enthalten folgende Gruppierungsniveaus:

- Ereignisse, die sich gleichzeitig ereignen, bilden eine einzelne Transaktion bzw. ein Element-Set.
- Jedes Element bzw. jedes Element-Set gehört zu einer Transaktionsgruppe. Beispiel: Ein gekaufter Artikel gehört zu einem Kunden, ein bestimmter Klick auf eine Seite gehört zu einem Web-Benutzer oder eine Komponente gehört zu einem produzierten Auto. Mehrere Element-Sets, die zu verschiedenen Zeiten auftreten und zur selben Transaktion gehören, bilden eine Sequenz.

### Optionen für ISW-Sequenzmodelle

Abbildung 5-18  
ISW-Sequenzknoten – Registerkarte "Modell"



**Modellname.** Sie können den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen (oder dem Modelltyp in Fällen, in denen kein solches Feld angegeben ist) generieren oder einen benutzerdefinierten Namen eingeben.

**Partitionierte Daten verwenden.** Wenn ein Partitionsfeld definiert ist, gewährleistet diese Option, dass nur Daten aus der Trainingspartition für die Modellerstellung verwendet werden. [Für weitere Informationen siehe Thema Partitionsknoten in Kapitel 4 in IBM SPSS Modeler 14.2- Quellen-, Prozess- und Ausgabeknoten.](#)

**Minimale Regelunterstützung (%).** Minimales Unterstützungsniveau für Assoziations- oder Sequenzregeln. Nur Regeln, die mindestens dieses Unterstützungsniveau erreichen, werden in das Modell eingeschlossen. Der Wert wird als  $A/B*100$  berechnet, wobei A die Anzahl der Gruppen ist, die alle in der Regel vorkommenden Elemente enthalten, und B ist die Gesamtzahl aller berücksichtigten Gruppen. Wenn Sie weitere gemeinsame Assoziationen oder Sequenzen wünschen, erhöhen Sie diese Einstellung.

**Minimale Regelkonfidenz (%).** Minimales Konfidenzniveau für Assoziations- oder Sequenzregeln. Nur Regeln, die mindestens dieses Konfidenzniveau erreichen, werden in das Modell eingeschlossen. Der Wert wird als  $m/n*100$  berechnet. Dabei ist  $m$  die Anzahl der Gruppen, die den per Join verbundenen Kopf der Regel (Sukzedens) und Regeltext (Antezedens) enthalten, und  $n$  ist die Anzahl der Gruppen, die den Regeltext enthalten. Wenn Sie zu viele oder uninteressante Assoziationen oder Sequenzen erhalten, sollten Sie diese Einstellung erhöhen. Wenn Sie zu wenige Assoziationen oder Sequenzen erhalten, sollten Sie diese Einstellung reduzieren.

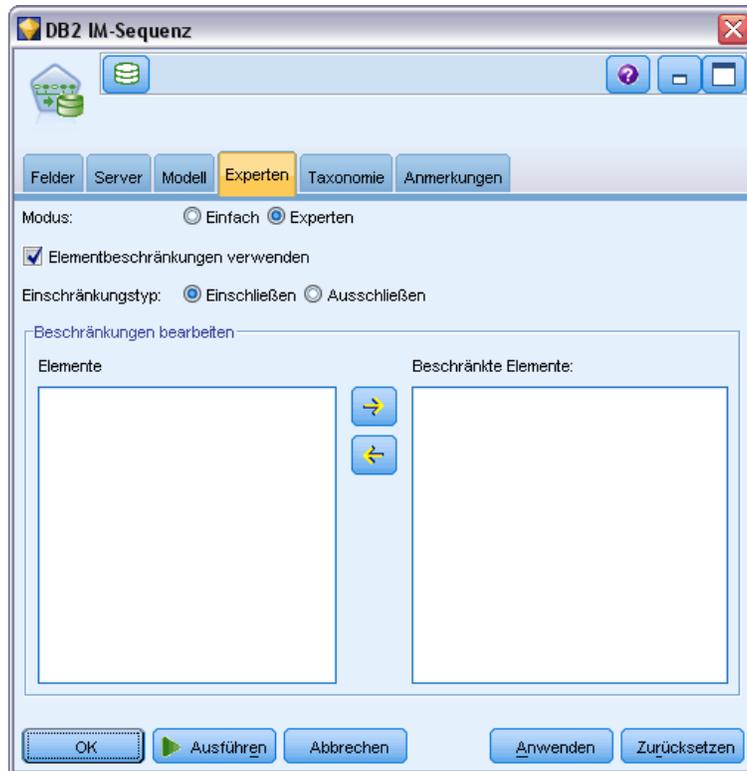
**Maximale Regelgröße.** Maximal zulässige Elemente in einer Regel, einschließlich des Sukzedens-Elements. Wenn die gewünschten Assoziationen oder Sequenzen relativ kurz sind, können Sie diese Einstellung reduzieren, um die Erstellung des Sets zu beschleunigen.

*Anmerkung:* Nur Knoten mit Transaktionseingabeformat werden gescort; Wahrheitstabellenformate (Tabellendaten) werden nicht verfeinert.

### ***Expertenoptionen für ISW-Sequenzen***

Sie können die Sequenzregeln angeben, die in den Ergebnissen ein- oder ausgeschlossen werden sollen. Wenn Sie bestimmte Elemente einschließen, werden die Regeln in das Modell übernommen, die mindestens eines der angegebenen Elemente enthalten. Wenn Sie angegebene Elemente ausschließen, werden die Regeln, die eines der angegebenen Elemente enthalten, in den Ergebnissen nicht berücksichtigt.

Abbildung 5-19  
ISW-Sequenzknoten – Registerkarte “Experten”



Wenn Elementbeschränkungen verwenden ausgewählt ist, werden alle Elemente, die Sie in die Beschränkungsliste eingefügt haben, abhängig von der Einstellung für den Einschränkungstyp, in den Ergebnissen ein- oder ausgeschlossen.

**Einschränkungstyp.** Wählen Sie, ob in den Ergebnissen diejenigen Assoziationsregeln ein- oder ausgeschlossen werden sollen, die die angegebenen Elemente enthalten.

**Beschränkungen bearbeiten.** Um ein Element zur Liste beschränkter Elemente hinzuzufügen, wählen Sie es in der Liste “Elemente” aus und klicken auf die Schaltfläche mit dem Rechtspfeil.

## ***ISW-Regression***

Der ISW-Regression-Knoten unterstützt die folgenden Regressionsalgorithmen:

- Transformation (Standard)
- Linear
- Polynomial
- RBF

### ***Transformationsregression***

Der ISW-Transformationsregressionsalgorithmus bildet Modelle als Entscheidungsbäume mit an den Baumblättern sitzenden Regressionsgleichungen. Das IBM Visualisierungs-Tool zeigt die Struktur dieser Modelle allerdings nicht an.

Der IBM® SPSS® Modeler-Browser zeigt die Einstellungen und Anmerkungen an. Die Modellstruktur kann jedoch nicht durchsucht werden. Es gibt relativ wenig durch den Benutzer konfigurierbare Aufbaueinstellungen.

### ***Lineare Regression***

Der lineare Regressionsalgorithmus für ISW geht von einer linearen Beziehung zwischen den erklärenden Feldern und dem Zielfeld aus. Er führt zu Modellen, die Gleichungen darstellen. Der vorhergesagte Wert weicht voraussichtlich vom beobachteten Wert ab, da eine Regressionsgleichung eine Näherung des Zielfelds darstellt. Die Differenz wird als "Rest" bezeichnet.

InfoSphere Warehouse Data Mining Modeling erkennt Felder, die keinen Erklärungswert aufweisen. Um zu bestimmen, ob ein Feld einen Erklärungswert aufweist, führt der lineare Regressionsalgorithmus zusätzlich zur autonomen Variablenauswahl statistische Tests durch. Wenn Sie die Felder kennen, die keinen Erklärungswert aufweisen, können Sie automatisch eine Untermenge der erklärenden Felder für kürzere Durchlaufzeiten auswählen.

Der lineare Regressionsalgorithmus bietet folgende Methoden zur automatischen Auswahl von Untermengen erklärender Felder:

**Schrittweise Regression.** Bei der schrittweisen Regression müssen Sie ein minimales Signifikanzniveau angeben. Nur Felder, deren Signifikanzniveau über dem angegebenen Wert liegt, werden vom linearen Regressionsalgorithmus verwendet.

**R-Quadrat-Regression** Die Methode der R-Quadrat-Regression identifiziert ein optimales Modell durch Optimierung eines Modellqualitätsmaßes. Es wird eines der folgenden Qualitätsmaße verwendet:

- Der quadrierte Korrelationskoeffizient nach Pearson
- Der korrigiert quadrierte Korrelationskoeffizient nach Pearson

Standardmäßig wählt der lineare Regressionsalgorithmus automatisch Untermengen erklärender Felder aus, indem er mithilfe des korrigierten quadrierten Korrelationskoeffizienten nach Pearson die Qualität des Modells optimiert.

### ***Polynomiale Regression***

Der Algorithmus für die polynomiale Regression bei ISW geht von einer polynominalen Beziehung aus. Ein polynomiales Regressionsmodell ist eine Gleichung, die aus folgenden Teilen besteht:

- Dem maximalen Grad der polynomen Regression
- Einer Näherung des Zielfelds
- Den erklärenden Feldern

### **RBF-Regression**

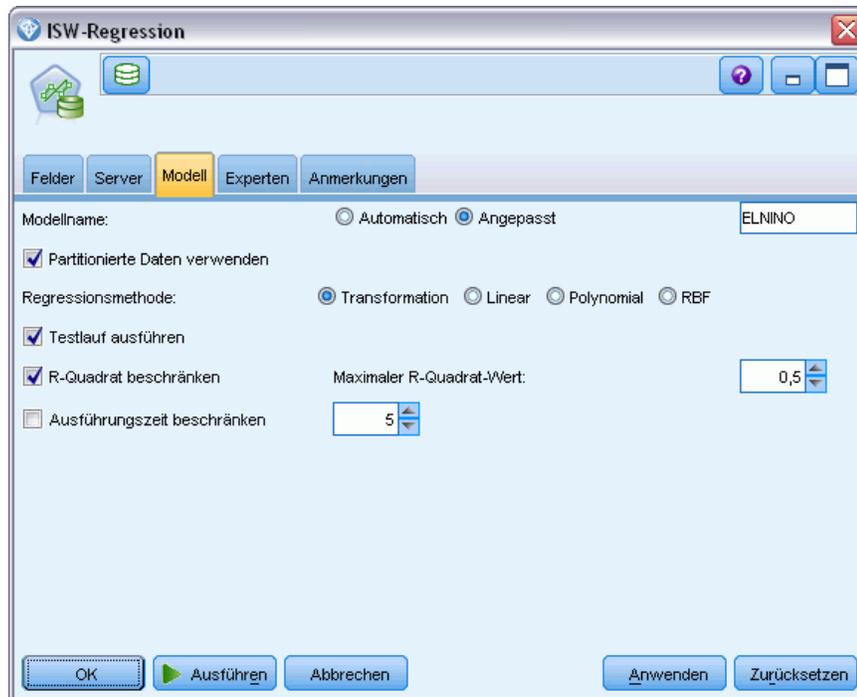
Der RBF-Regressionsalgorithmus für ISW geht von einer Beziehung zwischen den erklärenden Feldern und dem Zielfeld aus. Diese Beziehung kann als lineare Kombination Gauss'scher Funktionen ausgedrückt werden. Gauss'sche Funktionen sind spezifische radiale Basisfunktionen.

### **Optionen für ISW-Regressionsmodelle**

In der Registerkarte "Modell" des ISW-Regression-Knotens können Sie den Typ von Regressionsalgorithmus angeben, den Sie verwenden möchten, sowie:

- Ob partitionierte Daten verwendet werden sollen
- Ob ein Testlauf ausgeführt werden soll
- Eine Grenze für den  $R^2$ -Wert
- Eine Obergrenze für die Ausführungszeit

Abbildung 5-20  
ISW-Regressionknoten – Registerkarte "Modell"



**Partitionierte Daten verwenden.** Wenn ein Partitionsfeld definiert ist, gewährleistet diese Option, dass nur Daten aus der Trainingspartition für die Modellerstellung verwendet werden. [Für weitere Informationen siehe Thema Partitionsknoten in Kapitel 4 in IBM SPSS Modeler 14.2- Quellen-, Prozess- und Ausgabeknoten.](#)

**Regressionsmethode.** Wählen Sie den Regressionstyp, den Sie ausführen möchten. [Für weitere Informationen siehe Thema ISW-Regression auf S. 134.](#)

**Testlauf ausführen.** Sie können auswählen, dass ein Testlauf ausgeführt werden soll. Dann wird nach der Modellbildung in der Trainingspartition ein InfoSphere Warehouse Data Mining-Testlauf ausgeführt. Dabei wird ein Lauf über die Testpartition durchgeführt und es werden Modellqualitätsinformationen, Lift Charts etc. erzeugt.

**R-Quadrat beschränken.** Diese Option legt einen maximal tolerierten systematischen Fehler fest (den quadrierten Korrelationskoeffizienten nach Pearson,  $R^2$ ). Dieser Koeffizient misst die Korrelation zwischen dem Vorhersagefehler zu Verifizierungsdaten und den tatsächlichen Zielwerten. Er hat einen Wert zwischen 0 (keine Korrelation) und 1 (perfekte positive oder negative Korrelation). Der hier definierte Wert legt die Obergrenze für den akzeptablen systematischen Fehler des Modells.

**Ausführungszeit beschränken.** Geben Sie die gewünschte maximale Ausführungszeit in Minuten an.

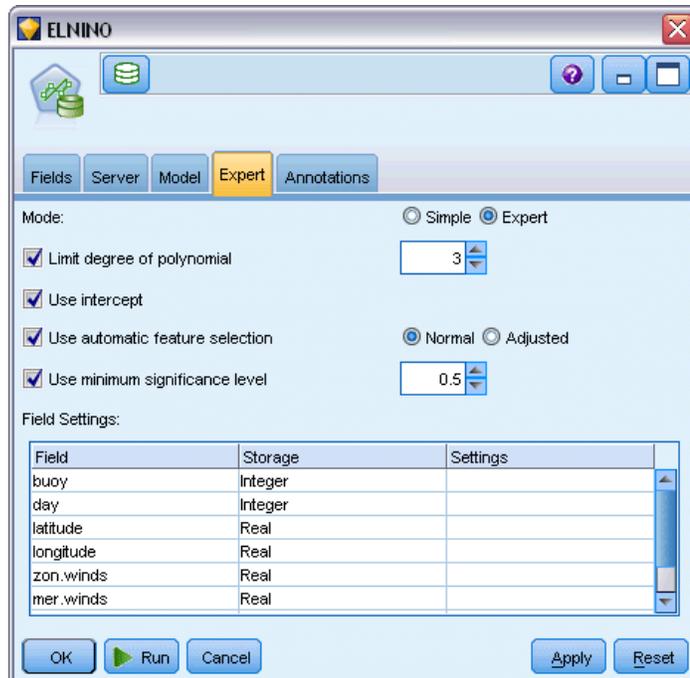
## Expertenoptionen für ISW-Regressionen

In der Registerkarte “Experten” des ISW-Regression-Knotens können Sie eine Reihe erweiterter Optionen für lineare, polynomiale oder RBF-Regression angeben.

### Expertenoptionen für lineare oder polynomiale Regression

Abbildung 5-21

Registerkarte “Experten” des ISW-Regressions-Knotens für lineare oder polynomiale Regression



**Grad des Polynoms begrenzen.** Legt den maximalen Grad der polynomen Regression fest. Wenn Sie den maximalen Grad der polynomen Regression auf 1 setzen, ist der Algorithmus für die polynomiale Regression identisch mit dem Algorithmus für die lineare Regression. Wenn

Sie einen hohen Wert als maximalen Grad der polynomialen Regression angeben, neigt der Algorithmus für die polynomiale Regression zur übermäßigen Anpassung. Dies bedeutet, dass das entstehende Modell zwar eine genaue Näherung der Trainingsdaten bietet, aber fehlschlägt, wenn es auf Daten angewendet wird, die nicht für das Training verwendet wurden.

**Konstanten Term verwenden.** Durch die Aktivierung dieser Einstellung wird erzwungen, dass die Regressionskurve durch den Ursprung verläuft. Dies bedeutet, dass das Modell keinen konstanten Term enthält.

**Automatische Merkmalsauswahl verwenden.** Wenn diese Einstellung aktiviert ist, versucht der Algorithmus eine optimale Untergruppe möglicher Prädiktoren zu bestimmen, wenn Sie kein minimales Signifikanzniveau angeben.

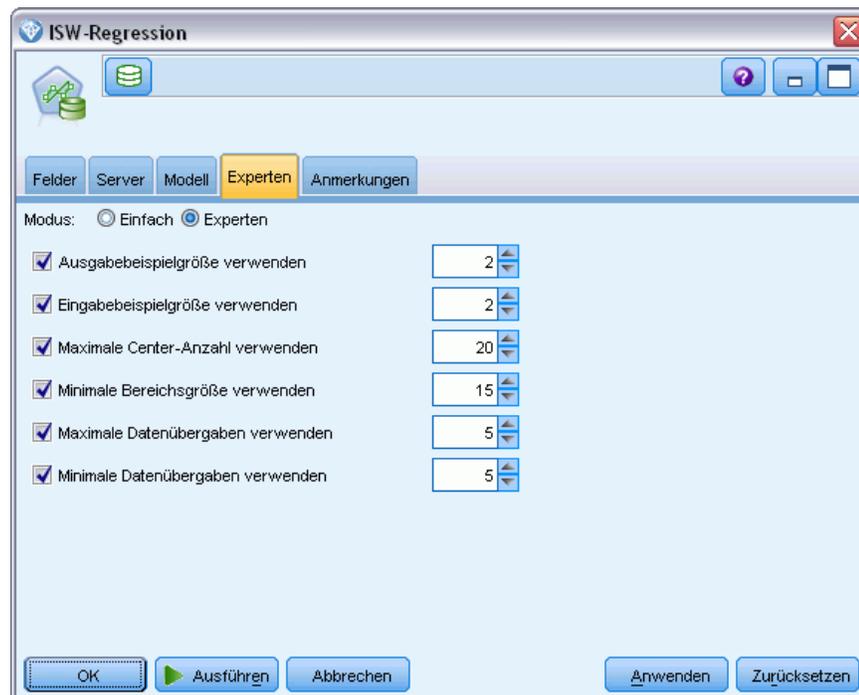
**Minimales Signifikanzniveau verwenden.** Wenn ein minimales Signifikanzniveau angegeben wurde, wird schrittweise Regression verwendet, um eine Untergruppe möglicher Prädiktoren zu bestimmen. Nur unabhängige Felder, deren Signifikanz über dem angegebenen Wert liegt, werden bei der Berechnung des Regressionsmodells berücksichtigt.

**Feldeinstellungen.** Zur Angabe von Optionen für einzelne Eingabefelder klicken Sie auf die entsprechende Zeile in der Spalte "Einstellungen" der Tabelle "Feldeinstellungen" und wählen <Einstellungen angeben>. [Für weitere Informationen siehe Thema Festlegen von Feldeinstellungen für Regression auf S. 139.](#)

### Expertenoptionen für RBF-Regression

Abbildung 5-22

Registerkarte "Experten" des ISW-Regressions-Knotens für RBF-Regression



**Stichprobenumfang für Ausgabe verwenden.** Definiert eine Stichprobe vom Typ “1-in-N” für die Modellverifikation und Tests.

**Stichprobenumfang für Eingabe verwenden.** Definiert eine Stichprobe vom Typ “1-in-N” für das Training.

**Maximale Zentrenanzahl verwenden.** Die maximale Anzahl von Zentren, die bei jedem Durchlauf erstellt werden. Da die Anzahl der Zentren sich während eines Durchlaufs verdoppeln kann, ist die tatsächliche Anzahl der Zentren unter Umständen höher als die Anzahl, die Sie festlegen.

**Minimale Bereichsgröße verwenden.** Die minimale Anzahl von Datensätzen, die einer Region zugeordnet sind.

**Maximale Datendurchgänge verwenden.** Die maximale Anzahl der Durchgänge durch die Eingabedaten durch den Algorithmus. Wenn dieser Wert angegeben wird, muss er größer oder gleich der minimalen Anzahl der Durchgänge sein.

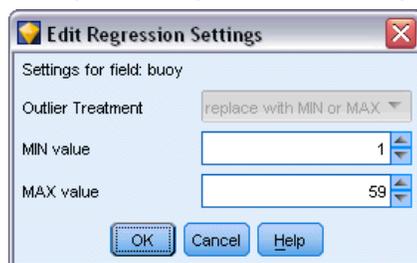
**Minimale Datendurchgänge verwenden.** Die minimale Anzahl der Durchgänge durch die Eingabedaten durch den Algorithmus. Geben Sie nur einen hohen Wert an, wenn Sie über ausreichend Trainingsdaten verfügen und sich sicher sind, dass ein gutes Modell vorhanden ist.

### ***Festlegen von Feldeinstellungen für Regression***

Hier können Sie den Wertebereich für ein einzelnes Eingabefeld angeben.

Abbildung 5-23

*Festlegen von Regressionseinstellungen für ein Eingabefeld*



**Min-Wert.** Zulässiger Mindestwert für dieses Eingabefeld.

**Max-Wert.** Zulässiger Höchstwert für dieses Eingabefeld.

## ***ISW Clustering***

Die Clustering-Mining-Funktion sucht in den Eingabedaten nach Merkmalen, die besonders häufig gemeinsam auftreten. Sie gruppiert die Eingabedaten in Cluster. Die Mitglieder der einzelnen Cluster weisen ähnliche Eigenschaften auf. Es gibt keine vorgefassten Vorstellungen davon, welche Muster in den Daten vorhanden sein könnten. Clustering ist ein Entdeckungsprozess.

Der ISW Clustering-Knoten bietet Ihnen die Wahl zwischen den folgenden Clustering-Methoden:

- demographic
- Kohonen

Die Technik des **demografischen Clustering**-Algorithmus arbeitet verteilungsbasiert. Verteilungsbasiertes Clustering bietet ein schnelles und natürliches Clustering sehr großer Datenbanken. Die Anzahl der Cluster wird automatisch ausgewählt (die maximale Clusterzahl kann angegeben werden). Eine Vielzahl von Parametern können durch den Benutzer konfiguriert werden.

Die Technik des **Kohonen-Clustering**-Algorithmus arbeitet zentrumsbasiert. Die Kohonen-Funktionskarte versucht, die Cluster-Zentren an Stellen zu setzen, die die Gesamtdistanz zwischen Datensätzen und Cluster-Zentren minimiert. Die Trennbarkeit von Clustern wird nicht berücksichtigt. Die Zentrumsvektoren werden auf einer Karte mit einer bestimmten Anzahl von Spalten und Zeilen angeordnet. Diese Vektoren sind miteinander verbunden, sodass nicht nur der Gewinnervektor, der am nächsten an einem Trainingsdatensatz liegt, angepasst wird, sondern auch die Vektoren in seiner Nachbarschaft. Je weiter die anderen Zentren jedoch entfernt sind, desto weniger werden sie angepasst.

## Modelloptionen für ISW Clustering

Auf der Registerkarte “Modell” des Clustering-Knotens können Sie die Methode zur Erstellung des Clusters sowie einige zugehörige Optionen festlegen.

Abbildung 5-24  
ISW-Knoten für Clustering – Registerkarte “Modell”



**Modellname.** Sie können den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen (oder dem Modelltyp in Fällen, in denen kein solches Feld angegeben ist) generieren oder einen benutzerdefinierten Namen eingeben.

**Partitionierte Daten verwenden.** Wenn ein Partitionsfeld definiert ist, gewährleistet diese Option, dass nur Daten aus der Trainingspartition für die Modellerstellung verwendet werden. [Für weitere Informationen siehe Thema Partitionsknoten in Kapitel 4 in IBM SPSS Modeler 14.2- Quellen-, Prozess- und Ausgabeknoten.](#)

**Cluster-Methode.** Wählen Sie die gewünschte Methode für die Cluster-Erstellung: Demografisch oder Kohonen. [Für weitere Informationen siehe Thema ISW Clustering auf S. 139.](#)

**Anzahl an Clustern beschränken.** Die Beschränkung der Anzahl der Cluster senkt die Ausführungszeit, da die Erstellung einer großen Zahl kleiner Cluster vermieden wird.

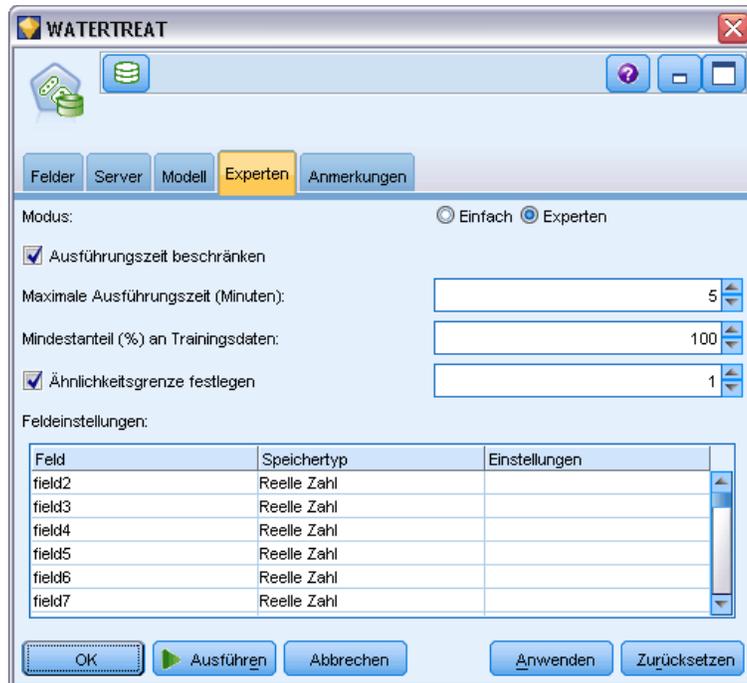
**Anzahl der Zeilen/Anzahl der Spalten.** (Nur Kohonen-Methode) Gibt die Anzahl der Zeilen und Spalten für die Kohonen-Funktionskarte an. (Nur verfügbar, wenn Anzahl der Kohonen-Durchgänge beschränken aktiviert und Anzahl an Clustern beschränken deaktiviert ist.)

**Maximale Anzahl der Durchgänge.** (Nur Kohonen--Methoden) Gibt die Anzahl der Durchgänge an, die der Kohonen-Clustering-Algorithmus während der Trainingsläufe in den Daten durchführt. Bei jedem Durchgang werden die Zentrumsvektoren angepasst, um den Gesamtabstand zwischen Cluster-Zentren und Datensätzen zu minimieren. Außerdem sinkt der Umfang, in dem die Vektoren angepasst werden. Beim ersten Durchgang sind die Anpassungen grob. Im letzten Durchgang, ist der Betrag, um den die Zentren angepasst werden, recht klein. Es werden nur geringfügige Anpassungen vorgenommen.

### ***Expertenoptionen für ISW Clustering***

Auf der Registerkarte “Experten” des Clustering-Knotens können Sie erweiterte Optionen wie Ähnlichkeitsgrenzen, maximale Ausführungszeiten und Feldgewichtungen festlegen.

Abbildung 5-25  
ISW-Knoten für Clustering – Registerkarte “Experten”



**Ausführungszeit beschränken.** Markieren Sie dieses Kontrollkästchen, um Optionen zu aktivieren, mit denen Sie die aufgebrauchte Zeit für die Erstellung des Modells steuern können. Sie können eine Dauer in Minuten und/oder einen prozentualen Mindestanteil der zu verarbeitenden Trainingsdaten angeben.

**Ähnlichkeitsschwellenwert angeben.** (Nur demografisches Clustering) Die Untergrenze für die Ähnlichkeit von zwei Datensätzen, die demselben Cluster angehören. Der Wert 0,25 beispielsweise bedeutet, dass Datensätze mit Werten, die zu 25 % ähnlich sind, wahrscheinlich demselben Cluster zugewiesen werden. Ein Wert von 1,0 bedeutet, dass Datensätze identisch sein müssen, damit sie im selben Cluster erscheinen.

**Feldeinstellungen.** Zur Angabe von Optionen für einzelne Eingabefelder klicken Sie auf die entsprechende Zeile in der Spalte “Einstellungen” der Tabelle “Feldeinstellungen” und wählen <Einstellungen angeben>.

### ***Festlegen von Feldeinstellungen für Clustering***

Hier können Sie Optionen für einzelne Eingabefelder angeben.

Abbildung 5-26  
Festlegen von Cluster-Einstellungen für ein Eingabefeld



**Feldgewichtung.** Ordnet dem Feld während des Modellerstellungsvorgangs mehr oder weniger Gewicht zu. Wenn Sie beispielsweise glauben, dass dieses Feld für das Modell weniger wichtig als andere Felder ist, verringern Sie die Feldgewichtung im Verhältnis zu den anderen Feldern.

**Wertegewichtung.** Weist bestimmten Werten dieses Felds mehr oder weniger Gewicht zu. Einige Feldwerte sind eventuell weniger üblich als andere. Die Koinzidenz von seltenen Werten in einem Feld ist eventuell weniger signifikant für ein Cluster als die Koinzidenz von häufigen Werten. Sie können eine der folgenden Methoden wählen, um Werte für dieses Feld zu gewichten (in jedem Fall haben seltene Werte ein höheres Gewicht, während häufige Werte ein geringeres Gewicht haben):

- **Logarithmisch.** Weist jedem Wert gemäß dem Logarithmus seiner Wahrscheinlichkeit in den Eingabedaten ein Gewicht zu.
- **Probabilistisch.** Weist jedem Wert gemäß seiner Wahrscheinlichkeit in den Eingabedaten ein Gewicht zu.

Für jede Methode können Sie auch eine Option unter mit Kompensation wählen, um die Wertegewichtung, die jedem Feld zugeordnet wurde, zu kompensieren. Wenn Sie die Wertegewichtung kompensieren, ist die Gesamtwichtigkeit des gewichteten Felds gleich dem eines ungewichteten Felds. Dies gilt unabhängig von der Anzahl der möglichen Werte. Kompensierte Gewichtung beeinflusst nur die relative Wichtigkeit von Koinzidenzen innerhalb des Sets von möglichen Werten.

**Ähnlichkeitsskala verwenden.** Markieren Sie dieses Kontrollkästchen, wenn Sie das Ähnlichkeitsniveau für ein Feld anhand einer Ähnlichkeitsskala steuern möchten. Sie geben die Ähnlichkeitsskala als absolute Zahl an. Diese Angabe wird nur für aktive numerische Felder berücksichtigt. Wenn Sie keine Ähnlichkeitsskala angeben, wird der Standardwert (die Hälfte der Standardabweichung) verwendet. Um eine größere Anzahl an Clustern zu erhalten, verringern Sie die mittlere Ähnlichkeit zwischen Cluster-Paaren durch kleinere Ähnlichkeitsskalen für numerische Felder.

**Sonderfallbehandlung.** Ausreißer sind Feldwerte, die außerhalb des für das Feld angegebenen Wertebereichs liegen, wie durch Min-Wert und Max-Wert definiert. Sie können wählen, wie Ausreißerwerte für dieses Feld behandelt werden sollen.

- Standardmäßig bedeutet keine, dass keine besondere Aktion für Ausreißerwerte ergriffen wird.
- Wenn Sie Durch MIN oder MAX ersetzen wählen, wird ein Feldwert kleiner als Min-Wert oder größer als Max-Wert wie passend durch die Werte von MIN oder MAX ersetzt. Sie können in diesem Fall die Werte von MIN und MAX festlegen.
- Wenn Sie Als fehlend behandeln wählen, werden Ausreißer als fehlende Werte betrachtet und nicht berücksichtigt. Sie können in diesem Fall die Werte von MIN und MAX festlegen.

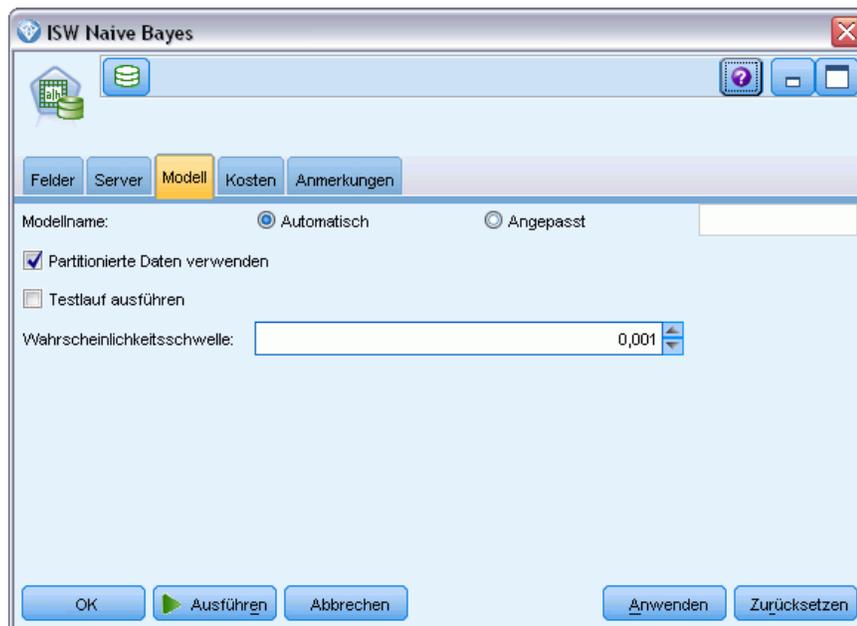
## ISW Naive Bayes

Naive Bayes ist ein bekannter Algorithmus für Klassifizierungsprobleme. Das Modell wird als *naiv* bezeichnet, weil es alle vorgeschlagenen Vorhersagevariablen als voneinander unabhängig behandelt. Naive Bayes ist ein schneller, skalierbarer Algorithmus, der für Kombinationen von Attributen und für das Zielattribut konditionale Wahrscheinlichkeiten berechnet. Aus den Trainingsdaten wird eine unabhängige Wahrscheinlichkeit ermittelt. Diese liefert die Wahrscheinlichkeit jeder Zielklasse anhand des Vorkommens der einzelnen Wertekategorien aus jeder einzelnen Eingabevariablen.

Der ISW-Naive Bayes-Klassifizierungsalgorithmus ist ein probabilistischer Klassifizierer. Er basiert auf Wahrscheinlichkeitsmodellen, die starke Unabhängigkeitsannahmen miteinbeziehen.

### Optionen für ISW Naive Bayes-Modelle

Abbildung 5-27  
ISW-Registerkarte "Modell" des Naive Bayes-Knotens



**Modellname.** Sie können den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen (oder dem Modelltyp in Fällen, in denen kein solches Feld angegeben ist) generieren oder einen benutzerdefinierten Namen eingeben.

**Partitionierte Daten verwenden.** Wenn ein Partitionsfeld definiert ist, gewährleistet diese Option, dass nur Daten aus der Trainingspartition für die Modellerstellung verwendet werden. [Für weitere Informationen siehe Thema Partitionsknoten in Kapitel 4 in IBM SPSS Modeler 14.2- Quellen-, Prozess- und Ausgabeknoten.](#)

**Testlauf ausführen.** Sie können auswählen, dass ein Testlauf ausgeführt werden soll. Dann wird nach der Modellbildung in der Trainingspartition ein InfoSphere Warehouse Data Mining-Testlauf ausgeführt. Dabei wird ein Lauf über die Testpartition durchgeführt und es werden Modellqualitätsinformationen, Lift Charts etc. erzeugt.

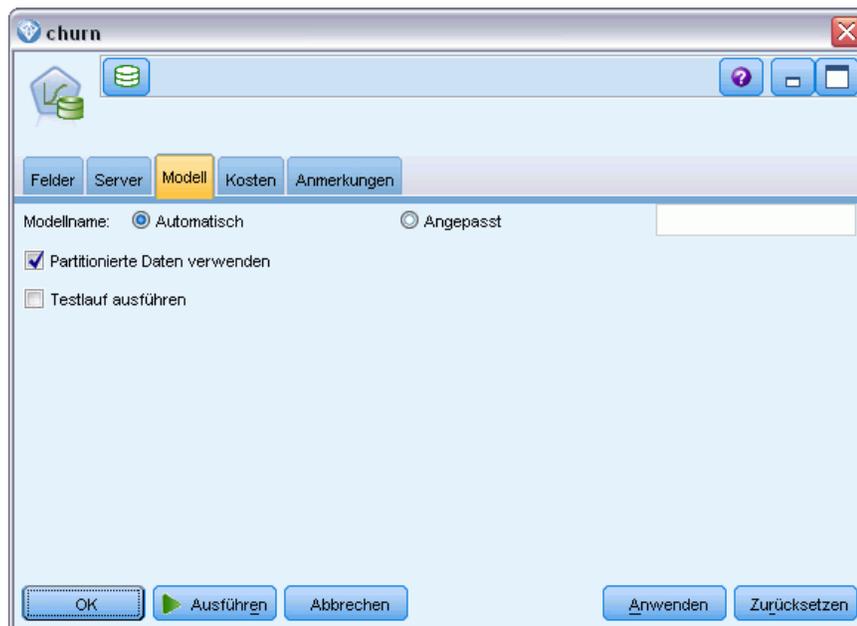
**Wahrscheinlichkeitsschwelle.** Die Wahrscheinlichkeitsschwelle definiert eine Wahrscheinlichkeit für alle Kombinationen der Prädiktor- und Zielwerte, die nicht in den Trainingsdaten erscheinen. Diese Wahrscheinlichkeit sollte zwischen 0 und 1 liegen. Der Standardwert ist 0,001.

## ISW Logistische Regression

Logistische Regression, auch als nominale Regression bekannt, ist ein statistisches Verfahren zur Klassifizierung von Datensätzen anhand der Werte der Eingabefelder. Sie verhält sich analog zur linearen Regression, aber der ISW-Logistische Regressions-Algorithmus bezieht sich nicht auf ein numerisches Zielfeld, sondern auf ein Flag-Zielfeld (binär).

### Optionen für logistische ISW-Regressionsmodelle

Abbildung 5-28  
ISW-Registerkarte "Modell" des logistischen Regressionsknotens



**Modellname.** Sie können den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen (oder dem Modelltyp in Fällen, in denen kein solches Feld angegeben ist) generieren oder einen benutzerdefinierten Namen eingeben.

**Partitionierte Daten verwenden.** Wenn ein Partitionsfeld definiert ist, gewährleistet diese Option, dass nur Daten aus der Trainingspartition für die Modellerstellung verwendet werden. [Für weitere Informationen siehe Thema Partitionsknoten in Kapitel 4 in IBM SPSS Modeler 14.2- Quellen-, Prozess- und Ausgabeknoten.](#)

**Testlauf ausführen.** Sie können auswählen, dass ein Testlauf ausgeführt werden soll. Dann wird nach der Modellbildung in der Trainingspartition ein InfoSphere Warehouse Data Mining-Testlauf ausgeführt. Dabei wird ein Lauf über die Testpartition durchgeführt und es werden Modellqualitätsinformationen, Lift Charts etc. erzeugt.

## ***ISW Time Series***

Mit dem ISW Time Series-Algorithmus können Sie zukünftige Ereignisse auf Basis bekannter vorangehender Ereignisse vorhersagen.

Ähnlich den üblichen Regressionsmethoden sagen Zeitreihenalgorithmen einen numerischen Wert vorher. Im Gegensatz zu den üblichen Regressionsmethoden konzentrieren sich Zeitreihenvorhersagen auf Zukunftswerte einer geordneten Zeitreihe. Diese Vorhersagen werden in der Regel als Prognosen bezeichnet.

Die Zeitreihenalgorithmen sind univariate Algorithmen. Das bedeutet, dass die unabhängige Variable eine Zeitspalte oder eine Reihenfolgenspalte ist. Die Prognosen basieren auf vorangehenden Werten. Sie basieren nicht auf anderen unabhängigen Spalten.

Zeitreihenalgorithmen unterscheiden sich von den üblichen Regressionsalgorithmen, weil sie nicht nur zukünftige Werte vorhersagen, sondern auch saisonale Zyklen in die Prognose miteinbeziehen.

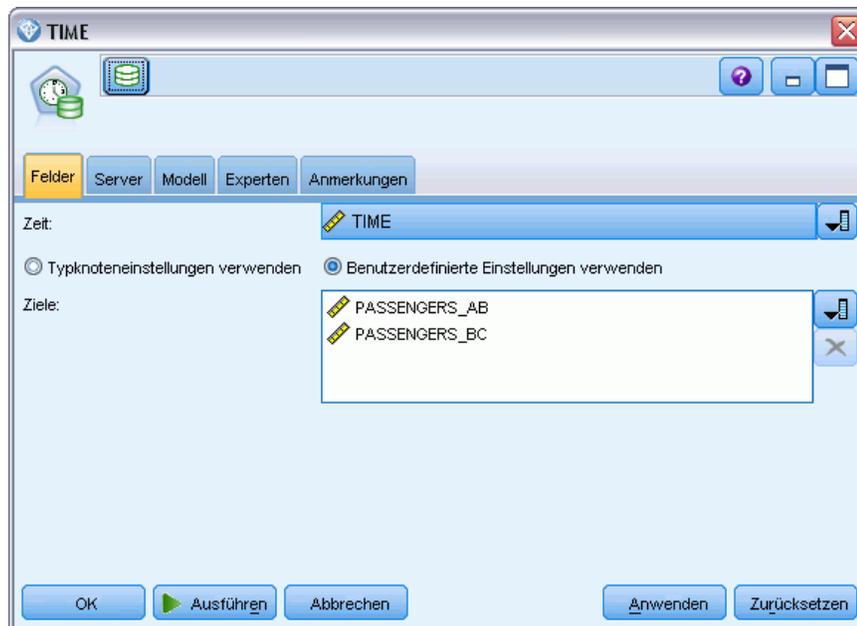
Die Mining-Funktion für Zeitreihen ermöglicht folgenden Algorithmen das Vorhersagen zukünftiger Trends:

- Autoregressiver integrierter gleitender Durchschnitt (AutoRegressive Integrated Moving Average, ARIMA).
- Exponentielles Glätten
- Saisonale Trend-Zerlegung

Welcher Algorithmus die besten Prognosen für Ihre Daten erstellt, hängt von unterschiedlichen Modellannahmen ab. Sie können alle Prognosen gleichzeitig berechnen. Die Algorithmen berechnen eine detaillierte Prognose einschließlich saisonalem Verhalten der ursprünglichen Zeitreihe. Wenn Sie den IBM InfoSphere Warehouse-Client installiert haben, können Sie das Zeitreihen-Visualisierungs-Tool verwenden und die erstellten Kurven vergleichen.

## ISW-Zeitreihen – Feldoptionen

Abbildung 5-29  
ISW-Zeitreihennoten – Registerkarte “Felder”



**Uhrzeit.** Wählen Sie das Eingabefeld aus, das die Zeitreihe enthält. Dieses Feld muss als Speichertyp “Datum”, “Zeit”, “Zeitstempel”, “Reelle Zahl” oder “Ganze Zahl” haben.

**Typknoteneinstellungen verwenden.** Diese Option weist den Knoten an, die Feldinformationen von einem weiter oben liegenden Typknoten zu verwenden. Dies ist die Standardeinstellung.

**Benutzerdefinierte Einstellungen verwenden.** Diese Option weist den Knoten an, die hier angegebenen Feldinformationen anstelle der in einem weiter oben liegenden Typknoten angegebenen zu verwenden. Geben Sie nach Auswahl dieser Option wie erforderlich die unten stehenden Felder an.

**Ziele.** Wählen Sie ein oder mehrere Zielfelder aus. Dies ist so, als würden Sie in einem Typknoten für die Rolle eines Felds den Wert *Ziel* festlegen.

## ISW-Zeitreihenmodelle – Optionen

Abbildung 5-30  
ISW-Zeitreihennoten – Registerkarte "Modell"



**Modellname.** Sie können den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen (oder dem Modelltyp in Fällen, in denen kein solches Feld angegeben ist) generieren oder einen benutzerdefinierten Namen eingeben.

**Algorithmen vorhersagen.** Wählen Sie die Algorithmen für die Modellierung aus. Sie können eine oder mehrere der folgenden Optionen auswählen:

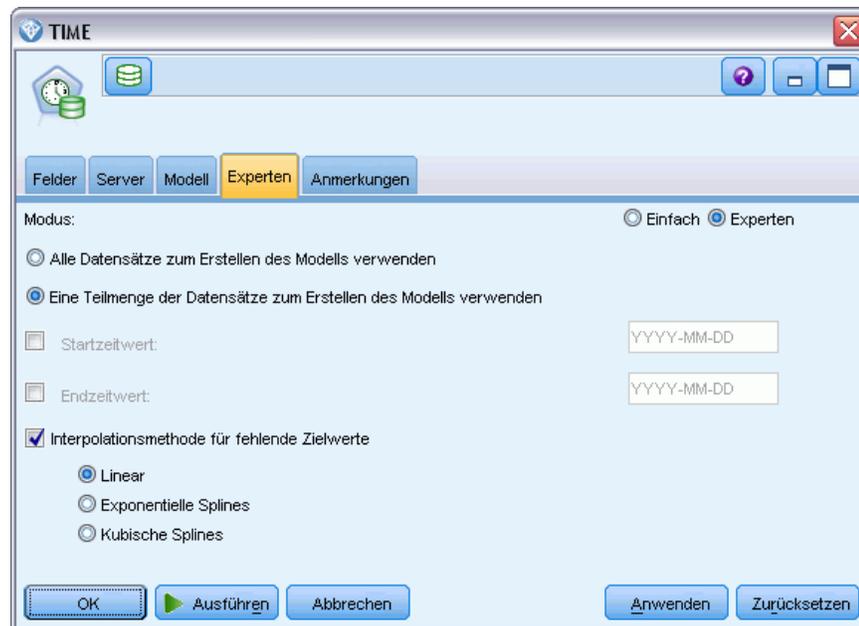
- ARIMA (X11 ARIMA)
- Exponentielles Glätten
- Saisonale Trend-Zerlegung.

**Endzeit der Prognose.** Geben Sie an, ob die Prognosenendzeit automatisch berechnet oder manuell angegeben werden soll.

**Zeitfeldwert.** Wenn die Endzeit der Prognose auf manuelle Eingabe eingestellt ist, geben Sie die Endzeit für die Prognose ein. Der Wert, den Sie eingeben können, hängt vom Feldtyp ab. Ist der Typ z. B. ganze Zahlen, die für Stunden stehen, können Sie 48 eingeben, sodass die Prognose endet, wenn die Daten für 48 Stunden verarbeitet wurden. Alternativ können Sie aufgefordert werden, ein Datum oder eine Uhrzeit als Endwert in das Feld einzugeben.

## ISW-Zeitreihen – Expertenoptionen

Abbildung 5-31  
ISW-Zeitreihenknoten – Registerkarte "Experten"



**Verwenden Sie alle Datensätze, um das Modell zu erstellen.** Dies ist die Standardeinstellung. Alle Datensätze werden analysiert, wenn das Modell erstellt wird.

**Verwenden Sie eine Teilmenge der Datensätze, um das Modell zu erstellen.** Wenn Sie das Modell nur aus einem Teil der verfügbaren Daten erstellen möchten, wählen Sie diese Option. Dies könnte z. B. notwendig sein, wenn übermäßig viele Wiederholungsdaten vorhanden sind.

Geben Sie Startzeitwert und Endzeitwert ein, um die zu verwendenden Daten zu identifizieren. Beachten Sie, dass die Werte, die Sie in diese Felder eingeben können, abhängig vom Zeitfeldtyp sind. Sie können z. B. die Anzahl der Stunden oder Tage, oder spezifische Daten oder Uhrzeiten angeben.

**Interpolierungsmethode für fehlende Zielwerte.** Wenn Sie Daten mit einem oder mehreren fehlenden Werten bearbeiten, wählen Sie aus, welche Methode zur Berechnung verwendet werden soll. Sie können eine der folgenden Optionen auswählen:

- Linear
- Exponentielle Splines
- Kubische Splines

## Anzeigen von ISW-Zeitreihenmodellen

ISW-Zeitreihenmodelle werden in der Form eines nicht verfeinerten Modells ausgegeben, das aus den Daten extrahierte Informationen enthält, aber nicht zum direkten Generieren von Vorhersagen geeignet ist.

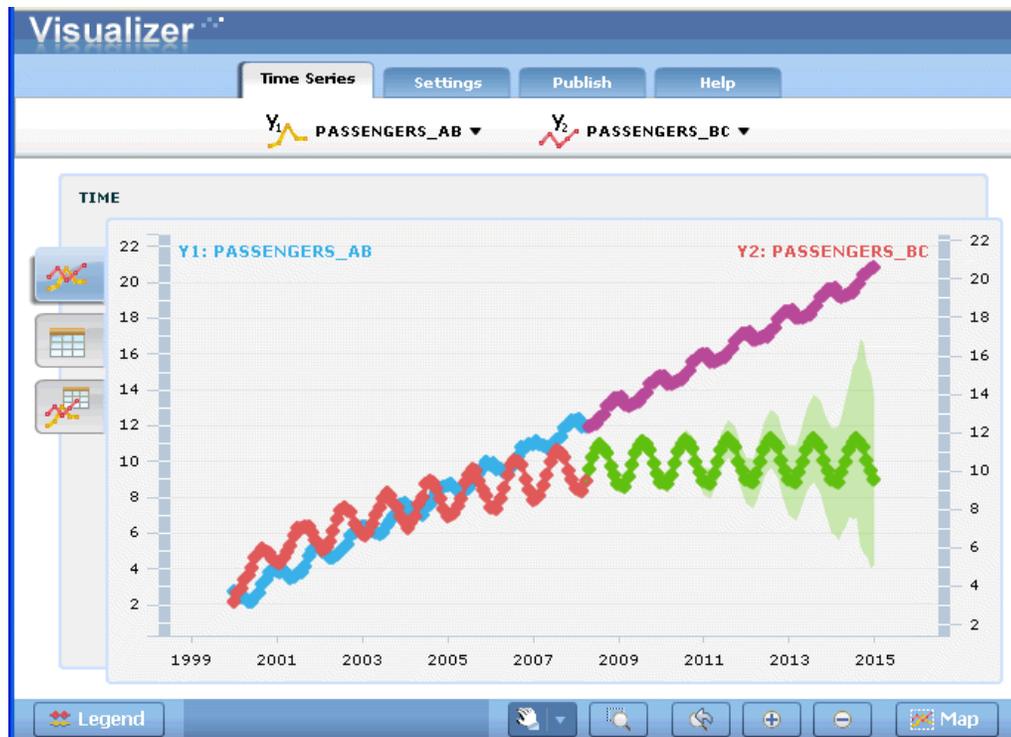
Abbildung 5-32  
Symbol für nicht verfeinerte Modelle



Für weitere Informationen siehe Thema Nicht verfeinerte Modelle in Kapitel 3 in *IBM SPSS Modeler 14.2-Modellierungsknoten*.

Wenn Sie den IBM InfoSphere Warehouse-Client installiert haben, können Sie das Zeitreihen-Visualisierungs-Tool verwenden, um eine grafische Darstellung Ihrer Zeitreihendaten zu erhalten.

Abbildung 5-33  
ISW-Zeitreihenmodell im Visualisierungs-Tool



So verwenden Sie das Zeitreihen-Visualisierungs-Tool:

- ▶ Stellen Sie sicher, dass Sie alle Aktionen für die Integration von IBM® SPSS® Modeler mit IBM InfoSphere Warehouse durchgeführt haben. Für weitere Informationen siehe Thema [Aktivieren der Integration mit InfoSphere Warehouse auf S. 107](#).
- ▶ Doppelklicken Sie auf das Symbol für nicht verfeinerte Modelle in der Modellpalette.

- Klicken Sie im Dialogfeld auf die Schaltfläche “Ansicht” auf der Registerkarte “Server”, um das Visualisierungs-Tool in Ihrem Standard-Web-Browser zu öffnen.

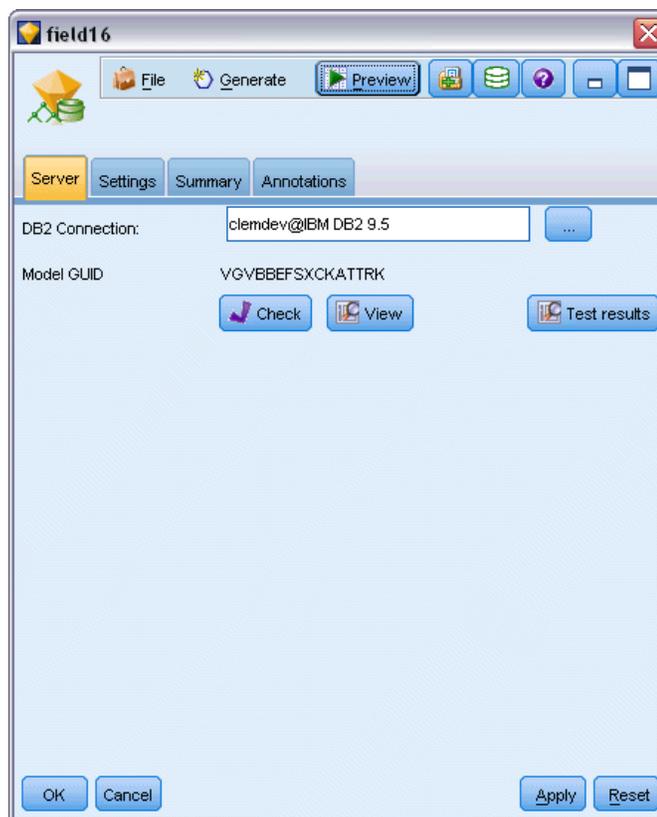
## ISW Data Mining Modell-Nuggets

Modelle können aus den in IBM® SPSS® Modeler enthaltenen Knoten für ISW-Entscheidungsbaum, -Assoziation, -Regression und -Clustering erstellt werden.

### ISW-Modell-Nugget – Registerkarte “Server”

Die Registerkarte “Server” bietet Optionen zur Durchführung von Konsistenzprüfungen und zum Starten des IBM Visualisierungs-Tools.

Abbildung 5-34  
ISW-Modell-Nugget – Registerkarte “Server”



IBM® SPSS® Modeler kann eine Konsistenzprüfung durchführen, indem sowohl im SPSS Modeler-Modell als auch im ISW-Modell eine identische, generierte Modellschlüsselzeichenkette gespeichert wird. Die Konsistenzprüfung erfolgt, wenn Sie auf der Registerkarte “Server” auf die Schaltfläche Überprüfen klicken. [Für weitere Informationen siehe Thema Verwalten von DB2-Modellen auf S. 117.](#)

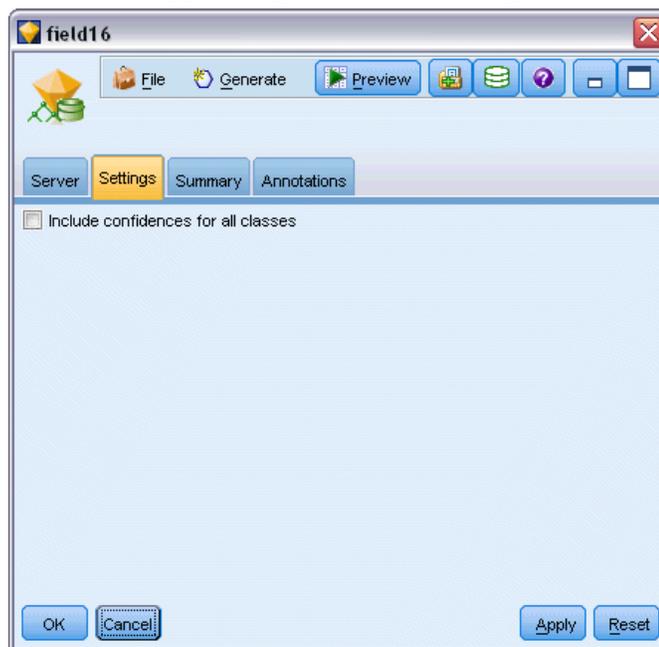
Das Visualisierungstool stellt die einzige Methode zum Durchsuchen von InfoSphere Warehouse Data Mining-Modellen dar. Das Tool kann optional mit InfoSphere Warehouse Data Mining installiert werden. [Für weitere Informationen siehe Thema Aktivieren der Integration mit InfoSphere Warehouse auf S. 107.](#)

- Klicken Sie auf Ansicht, um das Visualizer-Tool zu starten. Was das Tool anzeigt, hängt vom generierten Knotentyp ab. Beispielsweise gibt das Visualizer-Tool eine Ansicht der vorhergesagten Klassen aus, wenn es von einem ISW Entscheidungsbaum-Modell-Nugget gestartet wird.
- Klicken Sie auf Testergebnisse (nur Entscheidungsbäume und Sequenz), um das Visualizer-Tool zu starten und die Gesamtqualität des generierten Modells anzuzeigen.

### ***ISW-Modell-Nugget – Registerkarte “Einstellungen”***

In IBM® SPSS® Modeler wird in der Regel nur eine Vorhersage mit der zugehörigen Wahrscheinlichkeit oder Konfidenz erstellt. Zum Anzeigen der Wahrscheinlichkeiten jedes Ergebnisses (ähnlich wie die bei einer logistischen Regression) ist auf der Registerkarte “Einstellungen” des Modell-Nuggets zusätzlich eine Bewertungszeitoption verfügbar.

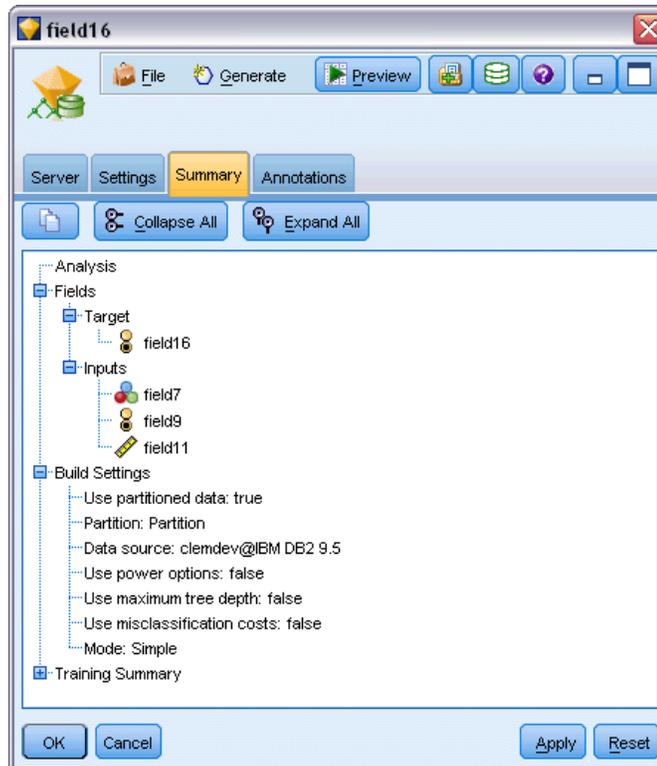
Abbildung 5-35  
*ISW-Modell-Nugget – Registerkarte “Einstellungen”*



**Konfidenzen für alle Klassen einschließen.** Für jedes der möglichen Ergebnisse für das Zielfeld wird eine Spalte hinzugefügt, die das Konfidenzniveau angibt.

## ISW-Modell-Nugget – Registerkarte “Übersicht”

Abbildung 5-36  
ISW-Modell-Nugget – Registerkarte “Übersicht”



Auf der Registerkarte “Übersicht” eines Modell-Nuggets finden Sie Informationen zum Modell selbst (*Analyse*), den im Modell verwendeten Feldern (*Felder*), den beim Erstellen des Modells verwendeten Einstellungen (*Aufbaueinstellungen*) und zum Modelltraining (*Trainingsübersicht*).

Beim ersten Durchsuchen des Knotens sind die Ergebnisse der Registerkarte “Übersicht” reduziert. Um die für Sie relevanten Ergebnisse anzuzeigen, vergrößern Sie sie mithilfe des Erweiterungssteuerelements links neben dem betreffenden Element oder klicken Sie auf die Schaltfläche Alles anzeigen, um alle Ergebnisse anzuzeigen. Um die Ergebnisse nach der Betrachtung wieder auszublenden, können Sie mit dem Erweiterungssteuerelement die gewünschten Ergebnisse reduzieren. Alternativ können Sie mit der Schaltfläche Alles ausblenden alle Ergebnisse ausblenden.

**Analyse.** Zeigt Informationen zum jeweiligen Modell an. Wenn Sie einen Analyseknoden ausgeführt haben, der an dieses Modell-Nugget angehängt ist, werden die Informationen aus dieser Analyse ebenfalls in diesem Abschnitt angezeigt. [Für weitere Informationen siehe Thema Analyseknoden in Kapitel 6 in IBM SPSS Modeler 14.2- Quellen-, Prozess- und Ausgabeknoten.](#)

**Felder.** Listet die Felder auf, die bei der Erstellung des Modells als Ziel bzw. als Eingaben verwendet werden.

**Aufbaueinstellungen.** Enthält Informationen zu den beim Aufbau des Modells verwendeten Einstellungen.

**Trainingsübersicht.** In diesem Abschnitt werden folgende Informationen angezeigt: der Typ des Modells, der für seine Erstellung verwendete Stream, der Benutzer, der ihn erstellt hat, der Erstellungszeitpunkt und die für den Aufbau des Modells benötigte Zeit.

## Beispiele für ISW Data Mining

IBM® SPSS® Modeler für Windows wird mit mehreren Demo-Streams geliefert, die den Vorgang des Database-Mining verdeutlichen. Diese Streams befinden sich im IBM® SPSS® Modeler-Installationsordner unter:

`\Demos\Database_Modeling\IBM DB2 ISW`

*Hinweis:* Dieser Demo-Ordner kann über die Programmgruppe “SPSS Modeler” im Windows-Startmenü aufgerufen werden.

Folgende Streams können nacheinander als Beispiel für den Database-Mining-Vorgang gestartet werden.

- `1_upload_data.str` — Bereinigt Daten und lädt sie aus einer Textdatei in DB2.
- `2_explore_data.str` — Beispiel für die Datenuntersuchung mit SPSS Modeler.
- `3_build_model.str` — Erstellt ein ISW-Entscheidungsbaummodell.
- `4_evaluate_model.str` — Beispiel für die Modellevaluation mit SPSS Modeler.
- `5_deploy_model.str` — Stellt das Modell für die datenbankinterne Bewertung bereit.

Die in den Beispiel-Streams verwendeten Datenmengen beziehen sich auf Kreditkartenanwendungen und stellen ein Klassifizierungsproblem mit einer Mischung aus kategorialen und stetigen Prädiktoren dar. Weitere Informationen über das Daten-Set enthält die folgende, mit SPSS Modeler installierte Datei unter:

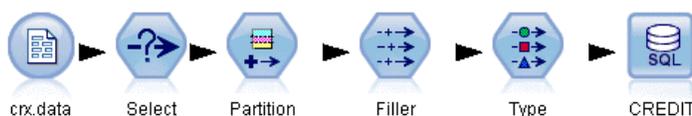
`\Demos\Database_Modeling\IBM DB2 ISW\crx.names`

Dieses Daten-Set befindet sich im UCI Machine Learning Repository unter <http://archive.ics.uci.edu/ml/>.

### Beispiel-Stream: Hochladen von Daten

Der erste Beispiel-Stream, `1_upload_data.str`, wird verwendet, um Daten aus einer Textdatei zu bereinigen und in DB2 zu laden.

Abbildung 5-37  
Beispiel-Stream zum Hochladen von Daten



Der Füllerknoten ist für die Behandlung von fehlenden Werten zuständig und ersetzt leere, aus der Textdatei `crx.data` eingelesene Felder durch `NULL`-Werte.

## Beispiel-Stream: Untersuchen von Daten

Der zweite Beispiel-Stream, *2\_explore\_data.str*, dient dazu, die Datenuntersuchung in IBM® SPSS® Modeler zu verdeutlichen.

Abbildung 5-38

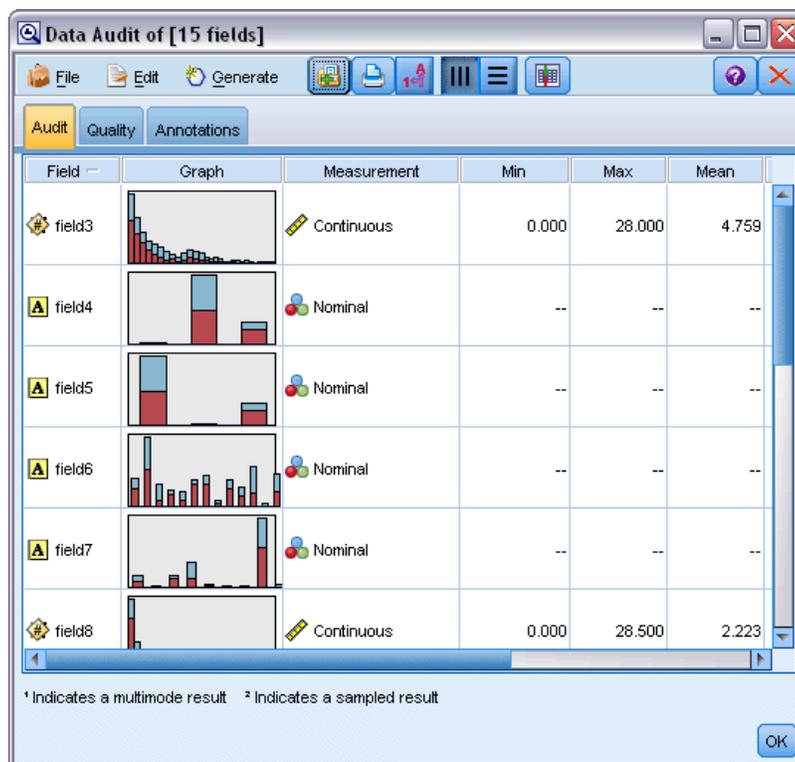
Beispiel-Stream zum Untersuchen von Daten



Ein typischerweise bei einer Datenuntersuchung durchgeführter Schritt besteht darin, die Daten mit einem Data Audit-Knoten zu verknüpfen. Der Data Audit-Knoten ist in der Palette “Ausgabeknoten” verfügbar.

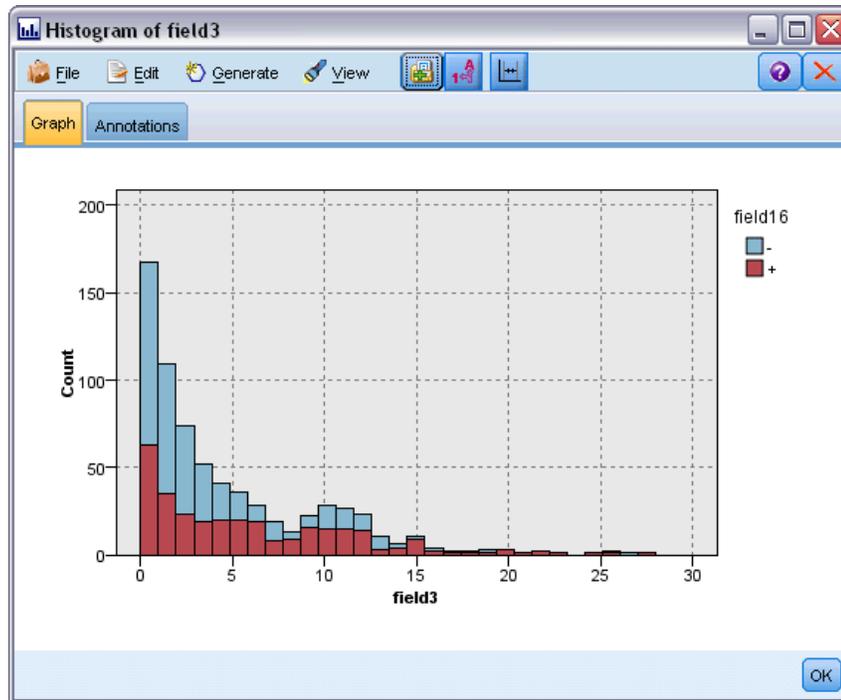
Abbildung 5-39

Data Audit-Ergebnisse



Mit der Ausgabe eines Data Audit-Knotens erhalten Sie einen allgemeinen Überblick über die Felder und die Datenverteilung. Wenn Sie im Fenster “Data Audit” auf ein Diagramm doppelklicken, wird ein detaillierteres Diagramm angezeigt, in dem Sie einzelne Felder eingehender untersuchen können.

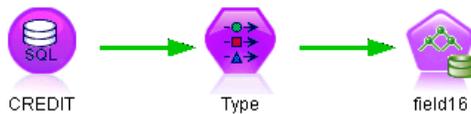
Abbildung 5-40  
Im Fenster "Data Audit" durch Doppelklicken auf dem Diagramm erzeugtes Histogramm



### Beispiel-Stream: Erstellen des Modells

Der dritte Beispiel-Stream, *3\_build\_model.str*, veranschaulicht die Modellerstellung in IBM® SPSS® Modeler. Sie können den Datenbankmodellknoten an den Stream anhängen und auf den Knoten doppelklicken, um Einstellungen für die Erstellung festzulegen.

Abbildung 5-41  
Beispiel-Stream für die Datenbank-Modellierung. Die violett eingezeichneten Knoten werden in der Datenbank ausgeführt.

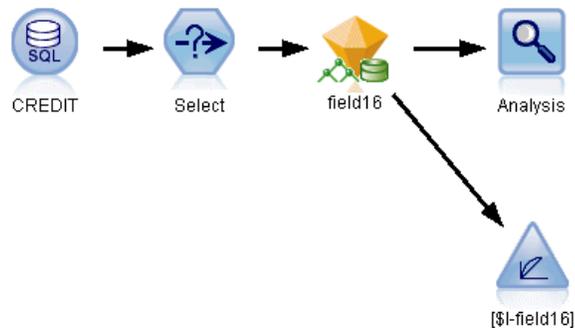


Über die Registerkarten "Modell" und "Experte" des Modellierungsknotens können Sie die maximale Baumtiefe anpassen und die weitere Aufteilung eines Knotens anhalten, von dem aus der ursprüngliche Entscheidungsbaum unter Angabe der maximalen Reinheit und der minimalen Fälle pro internen Knoten erstellt wurde. [Für weitere Informationen siehe Thema ISW-Entscheidungsbaum auf S. 123.](#)

## Beispiel-Stream: Evaluieren des Modells

Der vierte Beispiel-Stream, *4\_evaluate\_model.str*, veranschaulicht die Vorteile der Verwendung von IBM® SPSS® Modeler für die datenbankinterne Modellbildung. Sobald Sie das Modell ausgeführt haben, können Sie es wieder zu Ihrem Daten-Stream hinzufügen und das Modell mit verschiedenen von SPSS Modeler bereitgestellten Tools evaluieren.

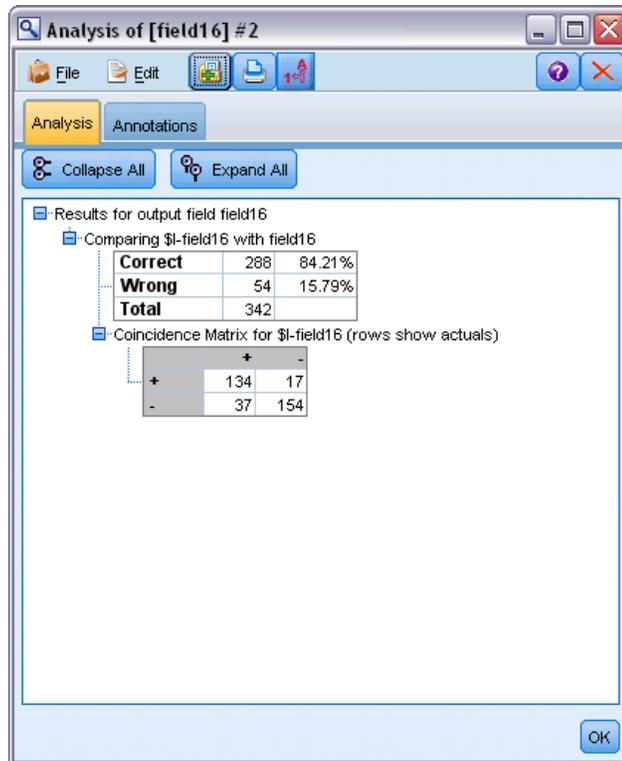
Abbildung 5-42  
Beispiel-Stream für die Modellevaluation



Beim ersten Öffnen des Streams ist das Modell-Nugget (*field16*) nicht im Stream enthalten. Öffnen Sie den CREDIT-Quellenknoten und stellen Sie sicher, dass Sie eine Datenquelle angegeben haben. Vorausgesetzt, Sie haben den Stream *3\_build\_model.str* ausgeführt, um ein *Feld 16*-Nugget in der Modelpalette zu erstellen, können Sie als Nächstes die getrennten Knoten ausführen, indem Sie auf die Schaltfläche Ausführen in der Symbolleiste klicken (die Schaltfläche mit einem grünen Dreieck). Dies startet ein Skript, das das *Feld 16*-Nugget in den Stream kopiert, es mit den vorhandenen Knoten verbindet und dann die Terminalknoten im Stream ausführt.

Sie können einen Analyseknotten (aus der Ausgabepalette) anfügen, um eine Fehlklassifizierungstabelle zu erstellen, aus der das Muster der Übereinstimmungen zwischen jedem generierten (vorhergesagten) Feld und dem zugehörigen Zielfeld ersichtlich wird. Führen Sie den Analyseknotten aus, um die Ergebnisse anzuzeigen.

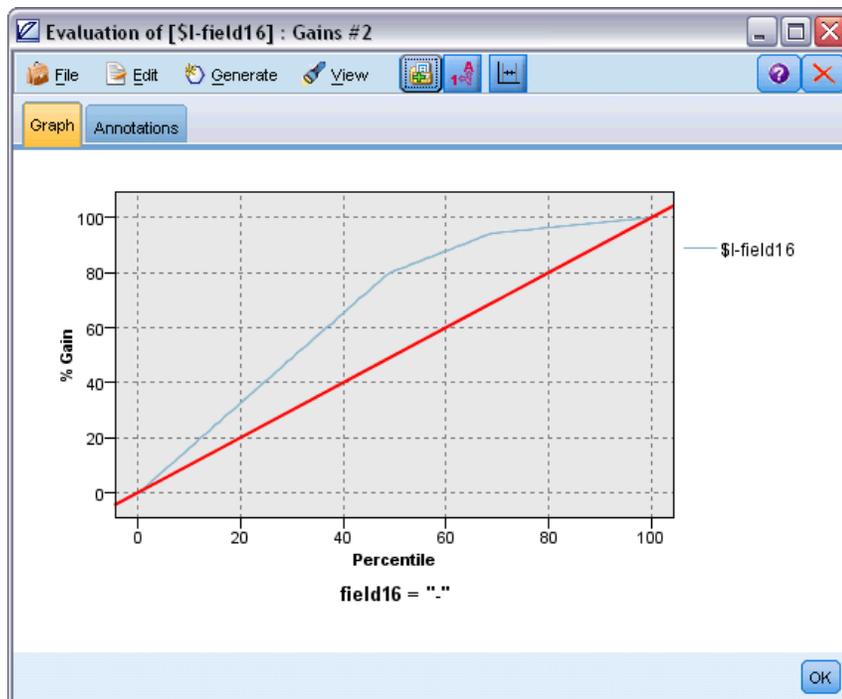
Abbildung 5-43  
Ergebnisse des Analyseknотens



Aus der generierten Tabelle geht hervor, dass 84,21 % der vom ISW-Entscheidungsstruktur-Algorithmus generierten Vorhersagen richtig sind.

Sie können auch ein Gewinn diagramm erstellen, das die Verbesserungen der Vorhersagegenauigkeit durch das Modell aufzeigt. Fügen Sie dem generierten Modell einen Evaluierungsknoten hinzu und führen Sie dann den Stream aus, um die Ergebnisse anzuzeigen.

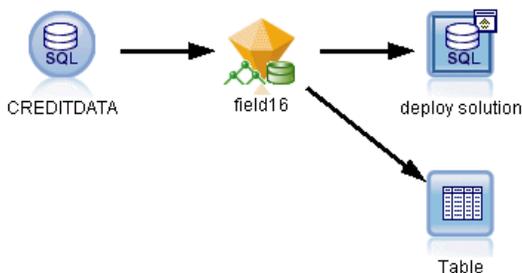
Abbildung 5-44  
Mit dem Evaluierungsknoten erstelltes Gewinn diagramm



### Beispiel-Stream: Bereitstellen des Modells

Sobald Sie mit der Genauigkeit des Modells zufrieden sind, können Sie es für die Verwendung mit externen Anwendungen oder für das Zurückschreiben der Scores in die Datenbank bereitstellen. Im Beispiel-Stream `5_deploy_model.str` werden die Daten aus der Tabelle CREDIT gelesen. Wenn der Datenbankexport-Knoten `deploy solution` ausgeführt wird, werden die Daten nicht wirklich gescort. Der Stream erzeugt stattdessen die PIM-Datei (Published Image) `credit_scorer.pim` und die PAR-Datei (Published Parameter) `credit_scorer.par`.

Abbildung 5-45  
Beispiel-Stream zum Bereitstellen des Modells



Wie im vorherigen Beispiel führt der Stream ein Skript aus, das das *Feld 16*-Nugget aus der Modellpalette in den Stream kopiert, es mit den vorhandenen Knoten verbindet und dann die Terminalknoten im Stream ausführt. In diesem Fall müssen Sie zuerst eine Datenquelle im Datenbankquellen- und Exportknoten angeben.

# **Datenbankmodellierung mit IBM Netezza Analytics**

## **IBM SPSS Modeler und IBM Netezza Analytics**

IBM® SPSS® Modeler unterstützt die Integration mit der Anwendung IBM® Netezza® Analytics, mit der erweiterte Analyseprozesse auf IBM Netezza-Servern ausgeführt werden können. Der Zugriff auf diese Funktionen erfolgt über die grafische Benutzeroberfläche und die am Workflow orientierte Entwicklungsumgebung von SPSS Modeler. So können Sie die Data-Mining-Algorithmen direkt in der IBM Netezza-Umgebung ausführen.

SPSS Modeler unterstützt die Integration der folgenden Algorithmen von Netezza Analytics.

- Decision Trees (Entscheidungsbäume)
- K-Means

Weitere Informationen zu den Algorithmen finden Sie im *Netezza Analytics-Entwicklerhandbuch* und *Netezza Analytics-Referenzhandbuch*.

## **Voraussetzungen für die Integration mit IBM Netezza Analytics**

Für die datenbankinterne Modellierung mit IBM® Netezza® Analytics gelten die folgenden Voraussetzungen. Wenden Sie sich ggf. an Ihren Datenbankverwalter, um sicherzustellen, dass diese Bedingungen erfüllt sind.

- IBM® SPSS® Modeler wird im lokalen Modus oder mit einer IBM® SPSS® Modeler Server-Installation unter Windows oder UNIX ausgeführt (ausgenommen zLinux, für das keine IBM Netezza-ODBC-Treiber zur Verfügung stehen).
- IBM Netezza Performance Server 6.0 oder höher, mit ausgeführtem IBM® SPSS® In-Database Analytics-Paket.
- Eine ODBC-Datenquelle zum Herstellen einer Verbindung mit einer IBM Netezza-Datenbank. [Für weitere Informationen siehe Thema Aktivieren der Integration mit IBM Netezza Analytics auf S. 162.](#)
- SQL-Erzeugung und -Optimierung aktiviert in SPSS Modeler. [Für weitere Informationen siehe Thema Aktivieren der Integration mit IBM Netezza Analytics auf S. 162.](#)

*Hinweis:* Für Datenbankmodellierung und SQL-Optimierung muss auf dem IBM® SPSS® Modeler-Computer die SPSS Modeler Server-Konnektivität aktiviert sein. Wenn diese Einstellung aktiviert ist, können Sie auf Datenbankalgorithmen zugreifen, SQL direkt aus SPSS Modeler per Pushback übertragen und auf SPSS Modeler Server zugreifen. Um den aktuellen Lizenzstatus zu überprüfen, klicken Sie im Menü "Hilfe" in SPSS Modeler auf Info. [Für weitere Informationen siehe Thema Verbindung mit IBM SPSS Modeler Server in Kapitel 3 in IBM SPSS Modeler 14.2- Benutzerhandbuch.](#)

## Aktivieren der Integration mit IBM Netezza Analytics

### Konfigurieren von IBM Netezza Analytics

Informationen zur Installation und Konfiguration von IBM® Netezza® Analytics finden Sie in der Netezza Analytics-Dokumentation—. Weitere Details enthält insbesondere das *Netezza Analytics-Installationshandbuch*—. Der Abschnitt zur *Einrichtung von Datenbankberechtigungen* in diesem Handbuch beinhaltet Details zu Skripts, die ausgeführt werden müssen, damit IBM® SPSS® Modeler-Streams in die Datenbank schreiben können.

### Erstellen einer ODBC-Datenquelle für IBM Netezza Analytics

Um die Verbindung zwischen der IBM Netezza-Datenbank und SPSS Modeler zu aktivieren, müssen Sie einen ODBC-System-DSN (Data Source Name, Datenquellennamen) erstellen.

Bevor Sie einen DSN erstellen, sollten Sie grundlegende Kenntnisse über ODBC-Datenquellen und -Treiber sowie über Datenbankunterstützung in SPSS Modeler besitzen. [Für weitere Informationen siehe Thema Datenzugriff in Kapitel 2 in IBM SPSS Modeler Server 14.2-Verwaltungs- und Leistungshandbuch.](#)

Wenn Sie mit IBM® SPSS® Modeler Server im verteilten Modus arbeiten, müssen Sie den DSN auf dem Server-Computer erzeugen. Wenn Sie im lokalen (Client-)Modus arbeiten, müssen Sie auf dem Client-Computer einen DSN erzeugen.

- ▶ Führen Sie von Ihrer *Netezza-Client-CD* aus die Datei *nzodbcsetup.exe* aus, um das Installationsprogramm zu starten. Folgen Sie den Anweisungen am Bildschirm, um den Treiber zu installieren. Vollständige Anweisungen finden Sie im *Installations- und Konfigurationshandbuch zu IBM Netezza ODBC, JDBC und OLE DB*.

- ▶ Erstellen Sie den DSN.

*Anmerkung:* Die Befehlsfolge ist abhängig von der jeweiligen Windows-Version.

- **Windows XP.** Wählen Sie im Menü “Start” die Option Systemsteuerung. Doppelklicken Sie auf Verwaltung und dann auf Datenquellen (ODBC).
- **Windows Vista** Wählen Sie im Menü “Start” die Option Systemsteuerung und dann Systemwartung. Doppelklicken Sie auf Verwaltung, wählen Sie dann Datenquellen (ODBC) und klicken Sie auf Öffnen.
- **Windows 7.** Wählen Sie im Menü “Start” die Option Systemsteuerung, dann System& Sicherheit und dann Verwaltung. Wählen Sie Datenquellen (ODBC) und klicken Sie dann auf Öffnen.
- ▶ Klicken Sie auf die Registerkarte System-DSN und dann auf Hinzufügen.
- ▶ Wählen Sie NetezzaSQL aus der Liste aus und klicken Sie auf Fertig stellen.
- ▶ Geben Sie auf der Registerkarte DSN Options (DSN-Optionen) des Bildschirms “Netezza ODBC Driver Setup” (Netezza ODBC-Treibereinrichtung) einen Datenquellennamen Ihrer Wahl, den Hostnamen oder die IP-Adresse des IBM Netezza-Servers, die Portnummer für die Verbindung, die Datenbank der verwendeten IBM Netezza-Instanz sowie den Benutzernamen und das Passwort für die Datenbankverbindung ein. Klicken Sie auf die Schaltfläche Hilfe, wenn Sie eine Erläuterung der Felder wünschen.

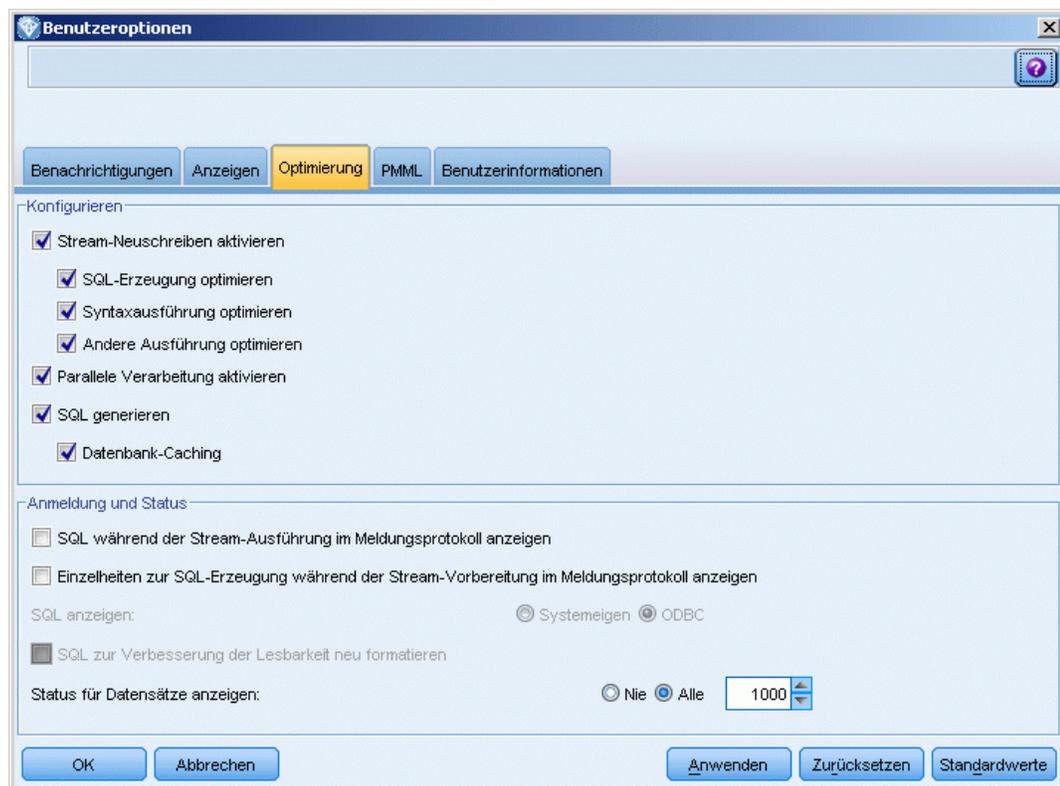
- ▶ Klicken Sie auf die Schaltfläche Test Connection (Verbindung testen) und stellen Sie sicher, dass Sie eine Verbindung zur Datenbank herstellen können.
- ▶ Klicken Sie mehrere Male auf OK, wenn Sie eine Verbindung hergestellt haben, um den Bildschirm "ODBC Data Source Administrator" (ODBC-Datenquellen-Administrator) zu schließen.

### **Aktivieren der SQL-Erzeugung und -Optimierung**

Da mit hoher Wahrscheinlichkeit mit großer Daten-Sets in gearbeitet wird, sollten Sie aus Leistungsgründen die SPSS Modeler-Optionen zur SQL-Erzeugung und -Optimierung aktivieren.

- ▶ Wählen Sie in den SPSS Modeler-Menüs folgende Befehlsfolge:  
Werkzeuge > Optionen > Benutzeroptionen

Abbildung 6-1  
Optimierungseinstellungen



- ▶ Klicken Sie dann auf die Registerkarte Optimierung.
- ▶ Überzeugen Sie sich, dass die Option SQL generieren aktiviert ist. Diese Einstellung ist für die Datenbank-Modellierung erforderlich.
- ▶ Wählen Sie SQL-Erzeugung optimieren und Andere Ausführung optimieren aus. (Diese Einstellungen sind nicht unbedingt erforderlich, werden aber zur Leistungsoptimierung empfohlen.)

Für weitere Informationen siehe Thema Festlegen von Optimierungsoptionen in Kapitel 12 in *IBM SPSS Modeler 14.2- Benutzerhandbuch*.

## **Erstellen von Modellen mit IBM Netezza Analytics**

Über die Registerkarte für die Datenbank-Modellierung in der Knotenpalette können Sie auf die IBM Netezza-Knoten zum In-Database Mining zugreifen.

### **Erläuterung der Daten**

Die Felder in der Datenquelle müssen entweder stetige oder kategoriale Variablen enthalten.

**Zielfeld.** Es kann nur ein einziges Feld als Ausgabefeld (Zielfeld) ausgewählt werden, in dem ein Ziel angegeben werden kann.

**Feld für Datensatz-ID.** Gibt das Feld an, über das jeder Fall eindeutig identifiziert wird. Dies kann beispielsweise ein ID-Feld sein, wie *CustomerID*.

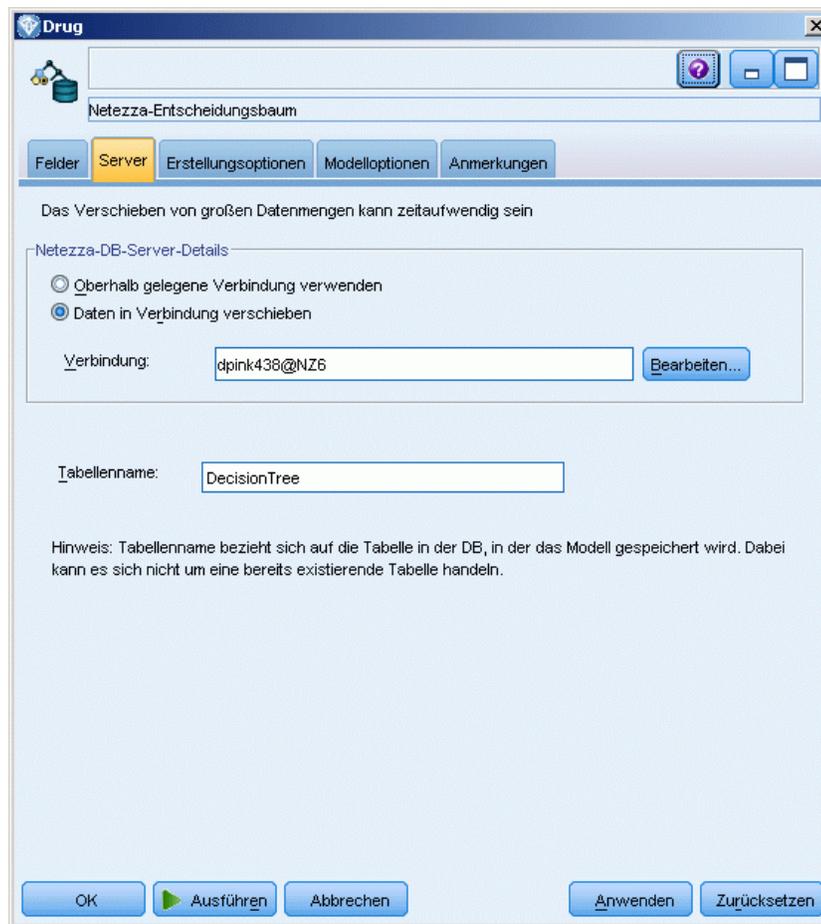
### **Allgemeine Kommentare**

- In IBM® SPSS® Collaboration and Deployment Services können keine Scoring-Konfigurationen mithilfe von Streams erstellt werden, die IBM Netezza-Datenbank-Modellierungsknoten enthalten.
- PMML-Export bzw. -Import ist für Modelle, die von den Netezza-Knoten erstellt wurden, nicht möglich.

## **Serveroptionen für Netezza-Modelle**

Auf dieser Registerkarte geben Sie die Verbindung zur IBM Netezza-Datenbank an, in der das Modell gespeichert werden soll.

Abbildung 6-2  
Netezza-Serveroptionen



**Netezza-DB-Server-Details.** Hier geben Sie die Verbindungsdetails für die für das Modell zu verwendende Datenbank an.

- **Oberhalb gelegene Verbindung verwenden.** (Standardeinstellung) Verwendet die Verbindungsdetails, die in einem oberhalb gelegenen Knoten, beispielsweise dem Datenbank-Quellenknoten, angegeben sind. *Hinweis:* Diese Option funktioniert nur, wenn alle oberhalb gelegenen Knoten SQL-Pushback verwenden können. In diesem Fall müssen die Daten nicht aus der Datenbank verschoben werden, da die SQL alle oberhalb gelegenen Knoten vollständig implementiert.
- **Daten in Verbindung verschieben.** Dient zum Verschieben der Daten in die hier angegebene Datenbank. Dadurch kann die Modellierung funktionieren, wenn sich die Daten in einer anderen IBM Netezza-Datenbank, einer Datenbank eines anderen Anbieters oder in einer Textdatei befinden. Darüber hinaus werden die Daten in die hier angegebene Datenbank zurückverschoben, wenn die Daten extrahiert wurden, da ein Knoten kein SQL-Pushback durchgeführt hat. Klicken Sie auf die Schaltfläche *Bearbeiten*, um eine Verbindung zu suchen und auszuwählen. *Vorsicht:* IBM® Netezza® Analytics wird in der Regel mit sehr großen Daten-Sets verwendet. Das Übertragen großer Datenmengen zwischen Datenbanken bzw. aus

einer Datenbank und wieder zurück kann sehr zeitaufwendig sein und sollte nach Möglichkeit vermieden werden.

**Tabellenname.** Der Name der Datenbanktabelle, in der das Modell gespeichert werden soll.

*Hinweis:* Dies muss eine neue Tabelle sein. Sie können für diesen Vorgang keine vorhandene Tabelle verwenden.

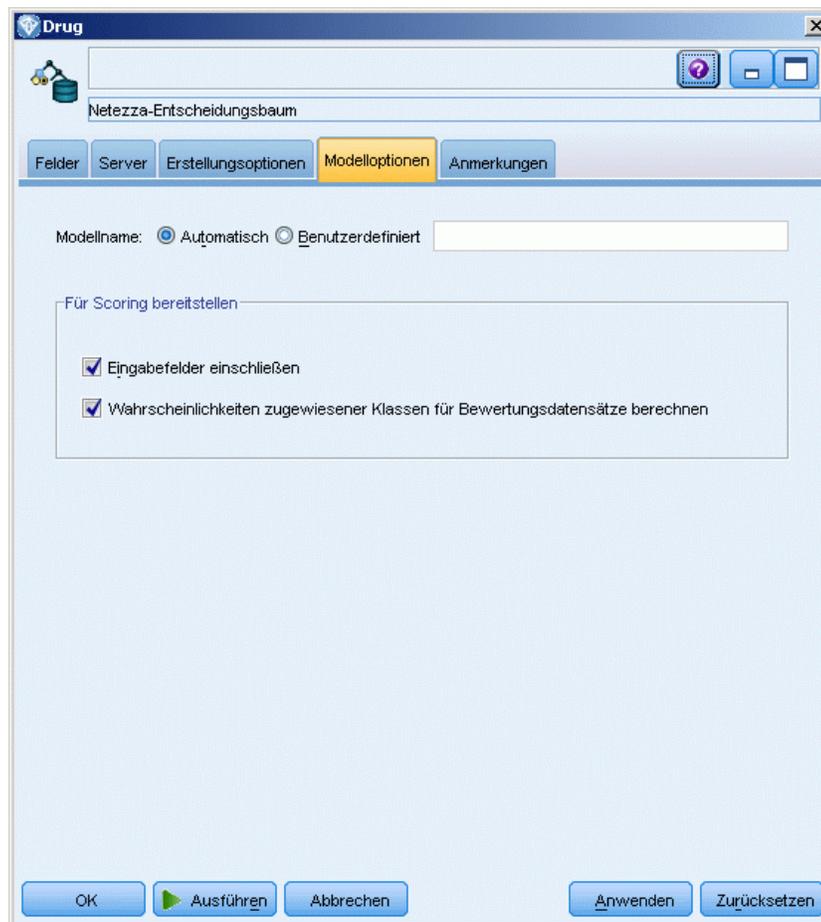
#### **Kommentare**

- Die für die Modellierung benutzte Verbindung muss nicht unbedingt mit der im Quellenknoten für einen Stream benutzen Verbindung identisch sein. Sie können beispielsweise einen Stream einsetzen, der auf die Daten einer IBM Netezza-Datenbank zugreift, die Daten für die Bereinigung und sonstige Bearbeitungen in IBM® SPSS® Modeler herunterlädt und dann zur Modellbildung in eine andere IBM Netezza-Datenbank lädt. Beachten jedoch, dass sich eine solche Konfiguration negativ auf die Leistung auswirken kann.
- Der Name der ODBC-Datenquelle wird in jeden SPSS Modeler-Stream eingebettet. Wenn ein auf einem Host erzeugter Stream auf einem anderen Host ausgeführt wird, muss der Name der Datenquelle auf beiden Hosts identisch sein. Alternativ kann für jeden Quellen- oder Modellierungsknoten auf der Registerkarte "Server" eine andere Datenquelle ausgewählt werden.

### **Modelloptionen für Netezza-Modelle**

Auf der Registerkarte "Modelloptionen" können Sie bestimmen, ob Sie einen Namen für das Modell angeben oder automatisch erstellen lassen möchten.

Abbildung 6-3  
Optionen für Entscheidungsbaummodelle



**Modellname.** Sie können den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen (oder dem Modelltyp in Fällen, in denen kein solches Feld angegeben ist) generieren oder einen benutzerdefinierten Namen eingeben.

**Für Scoring bereitstellen.** Sie können hier die Standardwerte für die Scoring-Optionen festlegen, die auf der Registerkarte "Einstellungen" des Modell-Nuggets angezeigt werden.

## ***Netezza-Entscheidungsbäume***

Ein Entscheidungsbaum ist eine hierarchische Struktur, die ein Klassifizierungsmodell darstellt. Mit einem Entscheidungsbaummodell können Sie ein Klassifizierungssystem entwickeln, die zukünftige Beobachtungen auf der Grundlage eines Satzes von Trainingsdaten vorhersagen oder klassifizieren. Die Klassifizierung hat die Form einer Baumstruktur, in der die Verzweigungen Teilungspunkte innerhalb der Klassifizierung darstellen. Die Teilungspunkte teilen die Daten rekursiv in Untergruppen auf, bis ein Endpunkt erreicht wird. Die Baumknoten an den Endpunkten werden als **Blätter** bezeichnet. Jedes Blatt weist den Mitgliedern seiner Untergruppe oder Klasse eine Beschriftung zu, die als **Klassenbeschriftung** bezeichnet wird.

## Instanzgewichtungen und Klassengewichtungen

Standardmäßig wird davon ausgegangen, dass alle Datensätze und Klassen die gleiche relative Wichtigkeit aufweisen. Sie können dies Ändern, indem Sie den Mitgliedern eines dieser Elemente bzw. beider Elemente individuelle Gewichte zuweisen. Dies kann beispielsweise dann sinnvoll sein, wenn die Datenpunkte in den Trainingsdaten nicht realistisch auf die verschiedenen Kategorien verteilt sind. Mit Gewichten können Sie das Modell verzerren, um einen Ausgleich für diejenigen Kategorien zu bewirken, die in den Daten unterrepräsentiert sind. Durch die Erhöhung des Gewichts für einen Zielwert sollte der Prozentsatz der richtigen Vorhersagen für die betreffende Kategorie erhöht werden.

Im Entscheidungsbaum-Modellierungsknoten können Sie zwei Arten von Gewichten angeben. **Instanzgewichtungen** weisen jeder Zeile von Eingabedaten ein Gewicht zu. Die gewichte werden üblicherweise für die meisten Fälle als 1,0 angegeben. Höhere oder niedrigere Werte erhalten nur diejenigen Fälle die wichtiger oder weniger wichtig sind als die Mehrheit, z. B.:

Datensatz-ID	Target	Instanzgewichtung
1	MedikamentA	1.1
2	MedikamentB	1.0
3	MedikamentA	1.0
4	MedikamentB	0.3

**Klassengewichtungen** weisen jeder Kategorie des Zielfelds ein Gewicht zu, z. B.:

Class	Klassengewichtung
MedikamentA	1.0
MedikamentB	1.5

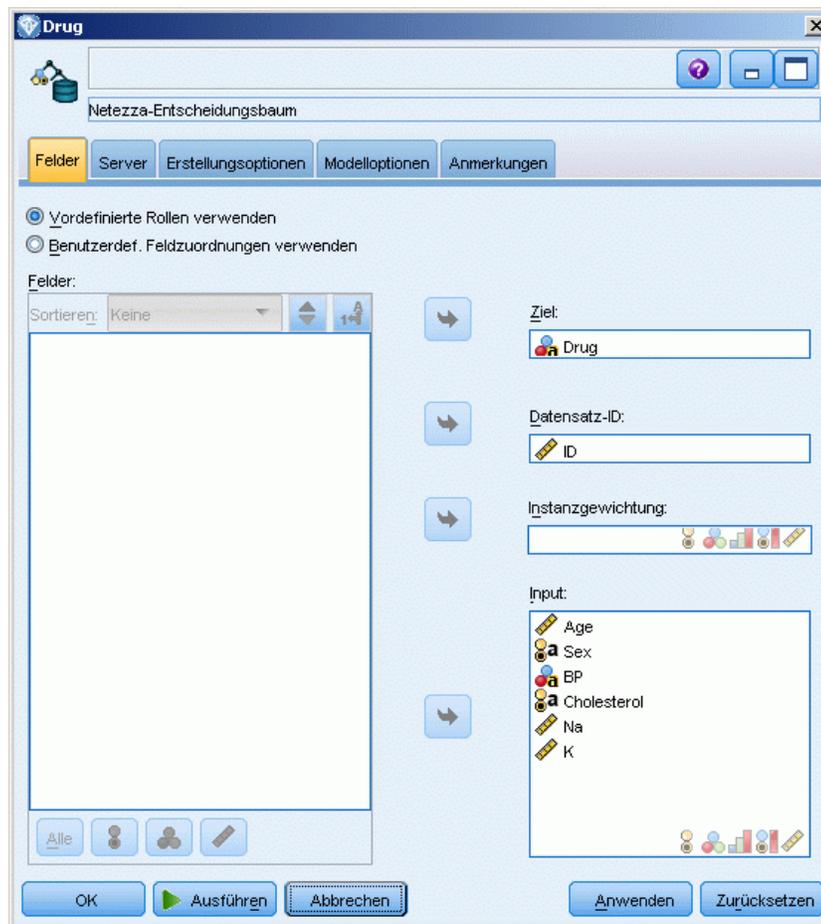
Beide Gewichtungstypen können gleichzeitig verwendet werden. In diesem Fall werden sie miteinander multipliziert und als Instanzgewichtungen verwendet. Wenn also die beiden vorigen Beispiele zusammen verwendet werden, führt dies zu folgenden Instanzgewichtungen beim Algorithmus.

Datensatz-ID	Berechnung	Instanzgewichtung
1	1.1*1.0	1.1
2	1.0*1.5	1.5
3	1.0*1.0	1.0
4	0.3*1.5	0.45

## Optionen für Netezza-Entscheidungsbaumfelder

Auf der Registerkarte "Felder" geben Sie an, ob Sie die Feldrolleneinstellungen verwenden möchten, die bereits in den Knoten oberhalb definiert wurden, oder ob Sie die Feldzuweisungen manuell vornehmen möchten.

Abbildung 6-4  
Optionen für Entscheidungsbaumfelder



**Vordefinierte Rollen verwenden** Bei dieser Option werden die Rolleneinstellungen (Ziele, Prädiktoren usw.) aus einem Typknoten oberhalb verwendet (oder die Registerkarte “Typen” eines weiter oben im Stream gelegenen Quellenknotens). [Für weitere Informationen siehe Thema Festlegen der Feldrolle in Kapitel 4 in IBM SPSS Modeler 14.2- Quellen-, Prozess- und Ausgabeknoten.](#)

**Benutzerdefinierte Feldzuweisungen verwenden** Wählen Sie diese Option, wenn Sie die Ziele, Prädiktoren und andere Rollen in diesem Bildschirm manuell zuweisen möchten.

**Felder.** Verwenden Sie die Pfeiltasten, um den verschiedenen Rollenfeldern rechts im Bildschirm Objekte aus dieser Liste manuell zuzuweisen. Die Symbole zeigen das gültige Messniveau für jedes Rollenfeld an.

Klicken Sie auf die Schaltfläche **Alle**, damit alle Felder in der Liste ausgewählt werden, oder klicken Sie auf die Schaltfläche für ein individuelles Messniveau, um alle Felder mit diesem Messniveau auszuwählen.

**Ziel.** Wählen Sie ein Feld als Ziel für die Vorhersage aus.

**Datensatz-ID.** Das Feld, das als eindeutiger Bezeichner für einen Datensatz verwendet wird. Die Werte in diesem Feld müssten für jeden Datensatz eindeutig sein (z. B. Kundennummern).

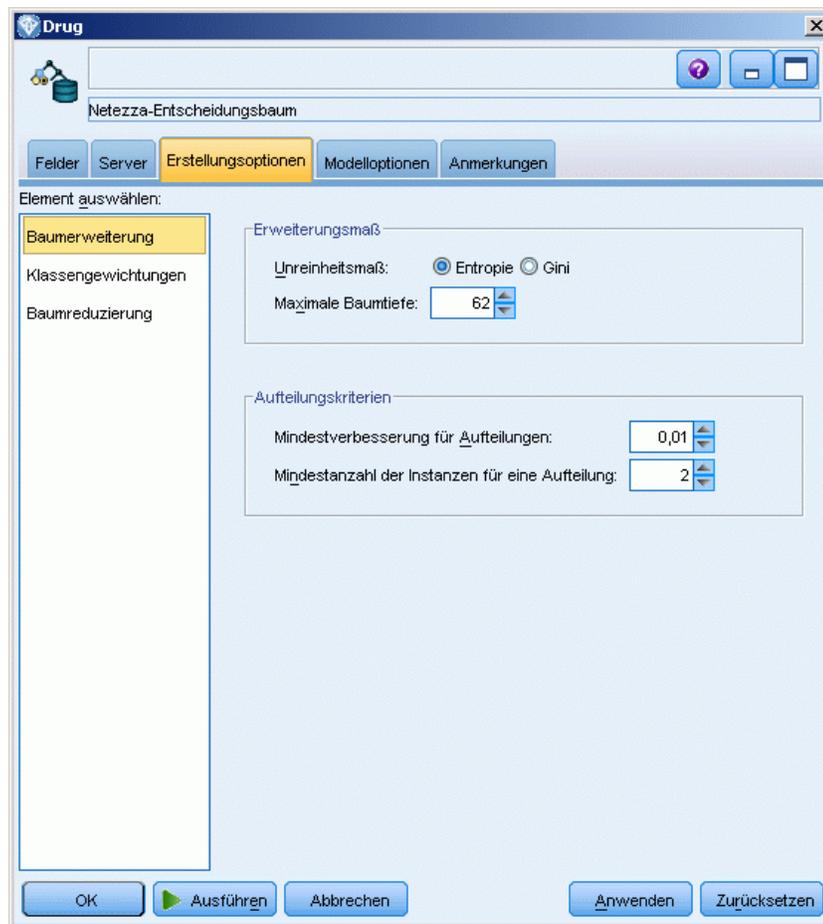
**Instanzgewichtung.** Indem Sie hier ein Feld angeben, können Sie anstelle der Klassengewichtungen oder zusätzlich zu den Klassengewichtungen (ein Gewicht pro Kategorie für das Zielfeld) Instanzgewichtungen (ein Gewicht pro Zeile an Eingabedaten) verwenden. Das hier angegebene Feld muss ein numerisches Gewicht für jede Zeile an Eingabedaten enthalten. [Für weitere Informationen siehe Thema Instanzgewichtungen und Klassengewichtungen auf S. 168.](#)

**Eingaben.** Wählen Sie das Eingabefeld bzw. die Eingabefelder aus. Dies ist so, als würden Sie in einem Typknoten für die Rolle eines Felds den Wert *Eingabe* festlegen.

### ***Erstellungsoptionen für Netezza-Entscheidungsbäume***

Auf der Registerkarte “Erstellungsoptionen” legen Sie sämtliche Optionen zum Erstellen des Modells fest. Selbstverständlich können Sie auch einfach auf die Schaltfläche Ausführen klicken, um ein Modell mit allen Standardoptionen zu erstellen, doch in der Regel werden Sie Einstellungen zur Anpassung an Ihre eigenen Zwecke vornehmen wollen.

Abbildung 6-5  
Entscheidungsbaum-Erstellungsoptionen für die Baumerweiterung



Sie können Erstellungsoptionen festlegen für:

- Baumerweiterung
- Gewichtungen für Klassenbeschriftungen
- Baumreduzierung

Die Optionen für die Baumerweiterung werden in diesem Abschnitt beschrieben.

**Erweiterungsmaß.** Diese Optionen steuern, wie die Baumerweiterung gemessen wird. Wenn Sie nicht die Standardwerte verwenden möchten, klicken Sie auf Anpassen und nehmen Sie die entsprechenden Änderungen vor.

- **Unreinheitsmaß.** Das Maß der Unreinheit, das verwendet wird, um die beste Position für eine Baumteilung zu ermitteln. **Unreinheit** bezieht sich auf das Ausmaß, in dem durch den Baum definierte Untergruppen in jeder Gruppe eine große Reihe von Ausgabefeldwerten besitzen.

Die Maße **Entropie** (Standard) und **Gini** werden unterstützt. Dies sind zwei häufig verwendete Unreinheitsmaße, die auf Wahrscheinlichkeiten der Zugehörigkeit zu einer Kategorie einer Verzweigung basieren.

- **Maximale Baumtiefe.** Die maximale Anzahl der Ebenen, auf die ein Baum unterhalb des Stammknotens erweitert werden kann (also die Anzahl der rekursiven Teilungen einer Stichprobe). Der Standardwert liegt bei 62. Dies ist die maximal mögliche Baumtiefe für Modellierungszwecke. Beachten Sie jedoch, dass im Viewer im Modell-Nugget maximal zehn Ebenen angezeigt werden können.

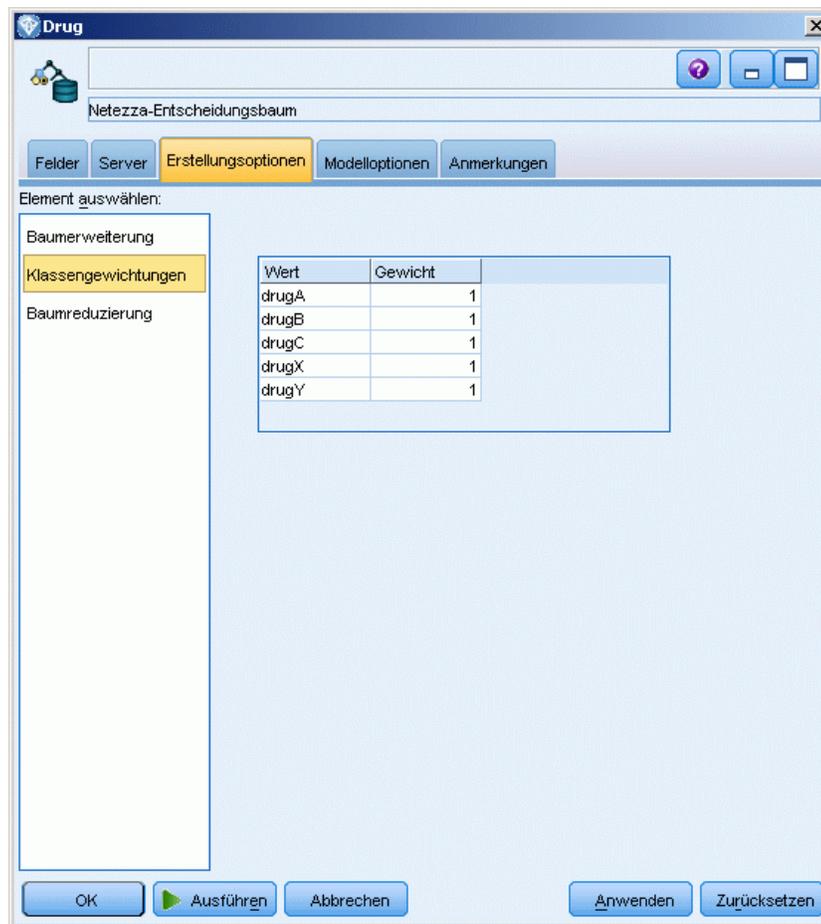
**Aufteilungskriterien.** Diese Optionen steuern, wann die Aufteilung des Baums aufhört. Wenn Sie nicht die Standardwerte verwenden möchten, klicken Sie auf Anpassen und nehmen Sie die entsprechenden Änderungen vor.

- **Mindestverbesserung für Aufteilungen.** Der Mindestwert der Unreinheitsreduzierung, bevor eine neue Aufteilung des Baums erfolgt. Das Ziel der Baumerstellung besteht darin, Untergruppen mit ähnlichen Ausgabewerte zu erstellen, also die Unreinheit in jedem Knoten zu minimieren. Wenn die beste für eine Verzweigung mögliche Aufteilung die Unreinheit um weniger als den durch die Aufteilungskriterien vorgegebenen Betrag reduziert, wird die Aufteilung nicht durchgeführt.
- **Mindestanzahl der Instanzen für eine Aufteilung.** Die Mindestanzahl von Datensätzen, die aufgeteilt werden kann. Wenn weniger nicht aufgeteilte Datensätze verbleiben, werden keine weiteren Aufteilungen durchgeführt. Sie können dieses Feld verwenden, um die Erstellung sehr kleiner Untergruppen im Baum zu verhindern.

### ***Netezza-Entscheidungsbaumknoten - Klassengewichtungen***

Hier können Sie den einzelnen Klassen Gewichte zuweisen. Standardmäßig wird allen Klassen der Wert 1 zugewiesen, sodass sie gleich gewichtet sind. Durch die Angabe von unterschiedlichen numerischen Gewichtungen für verschiedene Klassenbeschriftungen weisen Sie den Algorithmus an, die Trainings-Sets bestimmter Klassen entsprechend zu gewichten.

Abbildung 6-6  
Gewichtungsoptionen für Entscheidungsbaumklassen



Klicken Sie zum Ändern eines Gewichts in die Spalte Gewicht und nehmen Sie die gewünschten Änderungen vor.

**Wert:** Die Menge der Klassenbeschriftungen, abgeleitet aus den möglichen Werten des Zielfelds.

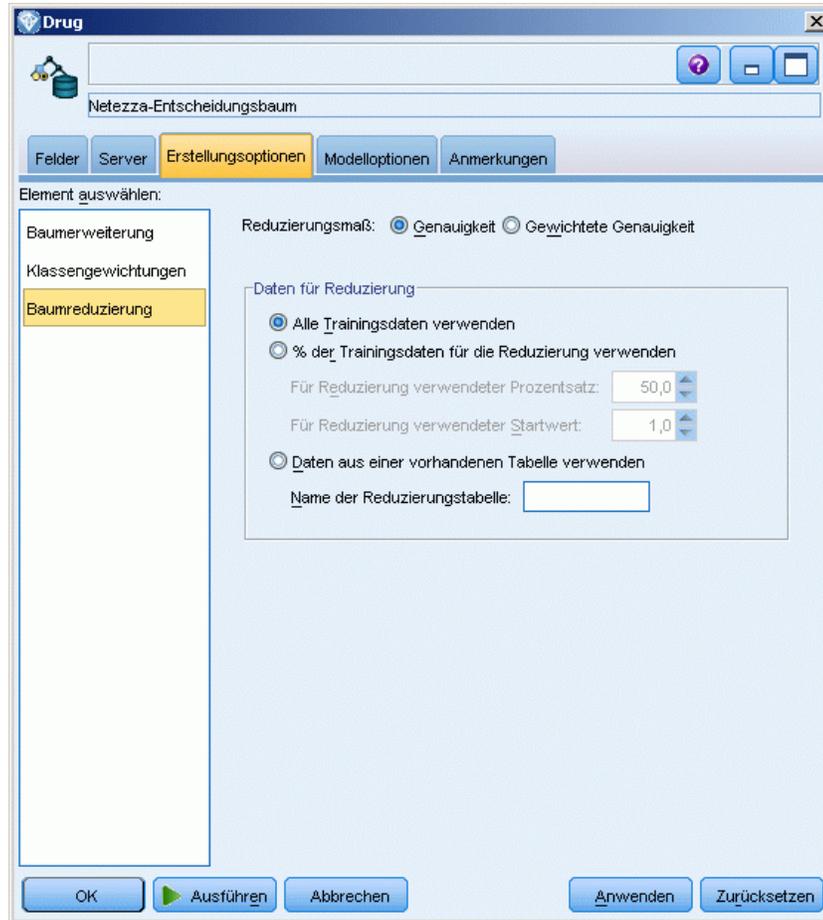
**Gewicht.** Die Gewichtung, die einer bestimmten Klasse zugewiesen werden soll. Durch Zuweisen einer höheren Gewichtung zu einer Klasse reagiert das Modell im Vergleich zu den anderen Klassen sensibler auf diese Klasse.

Klassengewichtungen können in Verbindung mit Instanzgewichtungen verwendet werden. [Für weitere Informationen siehe Thema Instanzgewichtungen und Klassengewichtungen auf S. 168.](#)

### Netezza-Entscheidungsbaumknoten - Reduzierungsoptionen

Sie können die Reduzierungsoptionen verwenden, um Reduzierungskriterien für den Entscheidungsbaum festzulegen. Ziel der Reduzierung ist es, das Risiko der übermäßigen Anpassung zu verringern, indem zu stark erweiterte Untergruppen entfernt werden, welche die erwartete Genauigkeit für neue Daten nicht verbessern.

Abbildung 6-7  
Optionen für die Entscheidungsbaumreduzierung



**Reduzierungsmaß.** Das standardmäßige Reduzierungsmaß, Genauigkeit, gewährleistet, dass die geschätzte Genauigkeit des Modells nach der Entfernung eines Blatts aus dem Baum innerhalb akzeptabler Grenzen bleibt. Nutzen Sie das Alternativmaß, Gewichtete Genauigkeit, wenn Sie die Klassengewichtungen in die Reduzierung mit einbeziehen möchten.

**Daten für die Reduzierung.** Sie können einen Teil oder alle Trainingsdaten verwenden, um die erwartete Genauigkeit der neuen Daten abzuschätzen. Alternativ können Sie zu diesem Zweck ein separates Daten-Set für die Reduzierung aus einer festgelegten Tabelle verwenden.

- **Alle Trainingsdaten verwenden.** Diese (standardmäßige) Option nutzt alle Trainingsdaten, um die Modellgenauigkeit zu schätzen.

- **% der Trainingsdaten für die Reduzierung verwenden.** Teilen Sie mithilfe dieser Option die Daten in zwei Gruppen (eine für das Training und eine für die Reduzierung) unter Verwendung des hier angegebenen Prozentsatzes für die Reduzierungsdaten. Im Feld Für Reduzierung verwendeter Startwert können Sie einen Zufallsstartwert angeben, um sicherzustellen, dass die Daten bei jeder Ausführung des Streams auf dieselbe Weise partitioniert werden.
- **Daten aus einer vorhandenen Tabelle verwenden.** Geben Sie den Tabellennamen eines separaten Daten-Sets für die Reduzierung an, anhand dessen die Modellgenauigkeit geschätzt wird. Diese Vorgehensweise wird als zuverlässiger betrachtet als die Nutzung von Trainingsdaten. Wenn Sie diese Option wählen, wird jedoch eventuell eine große Untermenge von Daten aus dem Trainings-Set entfernt, wodurch die Qualität des Entscheidungsbaum beeinträchtigt wird.

## ***Netezza-K-Means***

Der K-Means-Knoten implementiert den  $k$ -Means-Algorithmus, der eine Methode der Cluster-Analyse bietet. Mit diesem Knoten können Sie ein Clustering der Daten-Sets in einzelne Gruppen vornehmen.

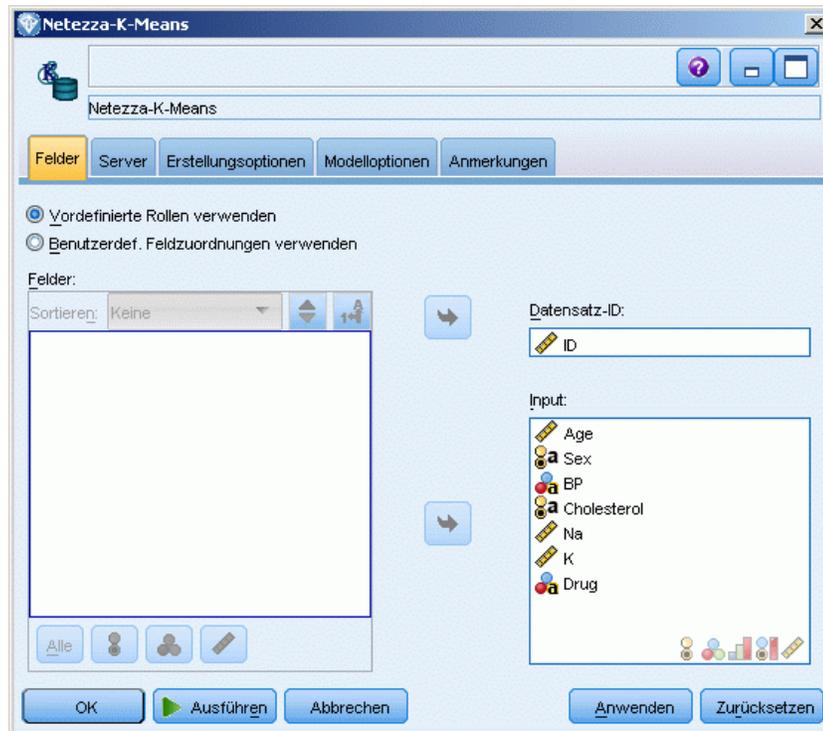
Bei dem Algorithmus handelt es sich um einen distanzbasierten Cluster-Algorithmus, der auf einer Distanzmetrik (Funktion) zur Messung der Ähnlichkeit zwischen Datenpunkten beruht. Die Datenpunkte werden dem nächsten Cluster gemäß der verwendeten Distanzmetrik zugewiesen.

Bei diesem Algorithmus werden mehrere Iterationen desselben Grundverfahren durchgeführt. Dabei wird jede Trainingsinstanz dem nächstgelegenen Cluster zugewiesen (in Bezug auf die angegebene Distanzfunktion, angewendet auf Instanz und Clusterzentrum). Anschließend werden alle Clusterzentren als Attribut-Mittelwertvektoren der Instanzen neu berechnet, die den jeweiligen Clustern zugewiesen wurden.

## ***K-Means-Feldoptionen von Netezza***

Auf der Registerkarte "Felder" geben Sie an, ob Sie die Feldrolleneinstellungen verwenden möchten, die bereits in den Knoten oberhalb definiert wurden, oder ob Sie die Feldzuweisungen manuell vornehmen möchten.

Abbildung 6-8  
K-Means-Feldoptionen



**Vordefinierte Rollen verwenden** Bei dieser Option werden die Rolleneinstellungen (Ziele, Prädiktoren usw.) aus einem Typknoten oberhalb verwendet (oder die Registerkarte “Typen” eines weiter oben im Stream gelegenen Quellenknotens). [Für weitere Informationen siehe Thema Festlegen der Feldrolle in Kapitel 4 in IBM SPSS Modeler 14.2- Quellen-, Prozess- und Ausgabeknoten.](#)

**Benutzerdefinierte Feldzuweisungen verwenden** Wählen Sie diese Option, wenn Sie die Ziele, Prädiktoren und andere Rollen in diesem Bildschirm manuell zuweisen möchten.

**Felder.** Verwenden Sie die Pfeiltasten, um den verschiedenen Rollenfeldern rechts im Bildschirm Objekte aus dieser Liste manuell zuzuweisen. Die Symbole zeigen das gültige Messniveau für jedes Rollenfeld an.

Klicken Sie auf die Schaltfläche **Alle**, damit alle Felder in der Liste ausgewählt werden, oder klicken Sie auf die Schaltfläche für ein individuelles Messniveau, um alle Felder mit diesem Messniveau auszuwählen.

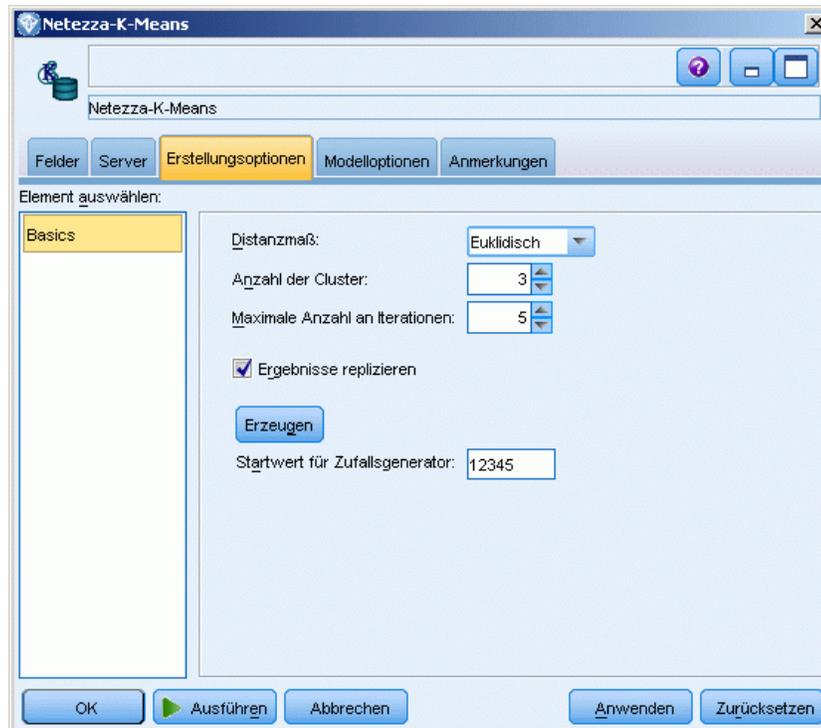
**Datensatz-ID.** Das Feld, das als eindeutiger Bezeichner für einen Datensatz verwendet wird.

**Eingaben.** Wählen Sie das Eingabefeld bzw. die Eingabefelder aus. Dies ist so, als würden Sie in einem Typknoten für die Rolle eines Felds den Wert *Eingabe* festlegen.

## K-Means-Erstellungsoptionen von Netezza

Auf der Registerkarte “Erstellungsoptionen” legen Sie sämtliche Optionen zum Erstellen des Modells fest. Selbstverständlich können Sie auch einfach auf die Schaltfläche Ausführen klicken, um ein Modell mit allen Standardoptionen zu erstellen, doch in der Regel werden Sie Einstellungen zur Anpassung an Ihre eigenen Zwecke vornehmen wollen.

Abbildung 6-9  
K-Means-Erstellungsoptionen



**Distanzmaß.** Die zur Messung des Abstands zwischen Datenpunkten zu verwendende Methode. Größere Abstände deuten auf größere Unähnlichkeiten hin. Folgende Optionen stehen zur Auswahl:

- **Euklidisch.** (Standardvorgabe). Der Abstand zwischen zwei Punkten wird anhand einer geradlinigen Verbindung zwischen en Punkten berechnet.
- **Manhattan.** Der Abstand zwischen zwei Punkten wird als Summe der absoluten Unterschiede zwischen ihren Koordinaten berechnet.
- **Canberra.** Ähnlich wie die Manhattan-Distanz, jedoch empfindlicher für Datenpunkte, die näher am Ursprung liegen.
- **Maximum.** Der Abstand zwischen zwei Punkten wird als der größte Unterschied in einer beliebigen Koordinatendimension berechnet.

**Anzahl an Clustern (k).** Geben Sie die Anzahl der zu erstellen Cluster an.

**Die maximale Anzahl von Iterationsschritten.** Bei diesem Algorithmus werden mehrere Iterationen desselben Vorgangs durchgeführt. Mit dieser Option können Sie das Modelltraining nach der angegebenen Anzahl von Iterationen beenden.

**Ergebnisse reproduzieren.** Aktivieren Sie dieses Kontrollkästchen, wenn Sie einen Zufallsstartwert festlegen möchten, mit dem Sie Analysen reproduzieren können. Sie können entweder eine ganze Zahl angeben oder auf Generieren klicken, wodurch eine pseudozufällige Ganzzahl erstellt wird.

## ***Verwalten von IBM Netezza Analytics-Modellen***

IBM® Netezza® Analytics-Modelle werden auf dieselbe Weise zum Zeichenbereich und zur Modellpalette hinzugefügt wie andere IBM® SPSS® Modeler-Modelle und können auf annähernd dieselbe Weise verwendet werden. Es gibt jedoch einige wichtige Unterschiede, die sich daraus ergeben, dass zurzeit jedes in SPSS Modeler erstellte Netezza Analytics-Modell auf ein in einem Datenbankserver gespeichertes Modell verweist. Damit ein Stream ordnungsgemäß funktioniert, muss eine Verbindung mit der Datenbank hergestellt werden, in der das Modell erstellt wurde, und die Modelltabelle darf nicht von einem externen Prozess geändert worden sein.

### ***Entscheidungsbaummodell-Nuggets von Netezza***

Das Entscheidungsbaummodell-Nugget zeigt die Ausgabe des Modellierungsvorgangs an und ermöglicht es Ihnen außerdem, einige Optionen für das Scoring des Modells festzulegen.

So können Sie die vollständige Modellausgabe für das Scoring anzeigen:

- ▶ Verbinden Sie einen Tabellenknoten mit dem Modell-Nugget.
- ▶ Öffnen Sie den Tabellenknoten.
- ▶ Klicken Sie auf Ausführen.

Blättern Sie im Tabellenausgabefenster nach rechts, um die Vorhersagen anzuzeigen. Im Feld *\$I-Feldname* werden die Prognosen angezeigt, während das Feld *\$IP-Feldname* (sofern vorhanden) die Konfidenzwerte für die einzelnen Prognosen enthält (von 0,0 bis 1.0).

### Netezza-Entscheidungsbaum-Nugget - Registerkarte "Modell"

Abbildung 6-10  
Ausgabe des Entscheidungsbaummodells

```

-- decision tree: "DecisionTree" --
K <= 0.055221
| K <= 0.037124
| | if true then class -> drugY
| | Na <= 0.678646
| | | BP = HIGH
| | | Na <= 0.656371
| | | | if true then class -> drugA
| | | | Age <= 41
| | | | | if true then class -> drugY
| | | | | if false then class -> drugB
| | | | K <= 0.040746
| | | | Age <= 43
| | | | | if true then class -> drugY
| | | | | if false then class -> drugX
| | | | BP = LOW
| | | | | Cholesterol = HIGH
| | | | | | if true then class -> drugC
| | | | | | if false then class -> drugX
| | | | | if false then class -> drugX
| | | K <= 0.05123
| | | | if true then class -> drugY
| | | | Na <= 0.760221
| | | | | if true then class -> drugX
| | | | | if false then class -> drugY
| | BP = HIGH
| | Age <= 50
  
```

Die Modellausgabe erfolgt in Form einer Textdarstellung des Entscheidungsbaums. Jede Zeile des Textes entspricht einem Knoten oder Blatt und die Einrückung steht für die Baumbene. Für einen Knoten wird die Aufteilungsbedingung angezeigt. Für ein Blatt wird die zugewiesene Klassenbeschriftung angezeigt.

### Netezza-Entscheidungsbaum-Nugget - Registerkarte "Einstellungen"

Auf der Registerkarte "Einstellungen" können Sie einige Optionen für das Scoring des Modells festlegen.

Abbildung 6-11  
Einstellungen für Netezza-Modelle



**Netezza-DB-Server-Details.** Hier geben Sie die Verbindungsdetails für die für das Modell zu verwendende Datenbank an.

- **Oberhalb gelegene Verbindung verwenden.** (Standardeinstellung) Verwendet die Verbindungsdetails, die in einem oberhalb gelegenen Knoten, beispielsweise dem Datenbank-Quellenknoten, angegeben sind. *Hinweis:* Diese Option funktioniert nur, wenn alle oberhalb gelegenen Knoten SQL-Pushback verwenden können. In diesem Fall müssen die Daten nicht aus der Datenbank verschoben werden, da die SQL alle oberhalb gelegenen Knoten vollständig implementiert.
- **Daten in Verbindung verschieben.** Dient zum Verschieben der Daten in die hier angegebene Datenbank. Dadurch kann die Modellierung funktionieren, wenn sich die Daten in einer anderen IBM Netezza-Datenbank, einer Datenbank eines anderen Anbieters oder in einer Textdatei befinden. Darüber hinaus werden die Daten in die hier angegebene Datenbank zurückverschoben, wenn die Daten extrahiert wurden, da ein Knoten kein SQL-Pushback durchgeführt hat. Klicken Sie auf die Schaltfläche *Bearbeiten*, um eine Verbindung zu suchen und auszuwählen. *Vorsicht:* IBM® Netezza® Analytics wird in der Regel mit sehr großen Daten-Sets verwendet. Das Übertragen großer Datenmengen zwischen Datenbanken bzw. aus einer Datenbank und wieder zurück kann sehr zeitaufwendig sein und sollte nach Möglichkeit vermieden werden.

**Modellname.** Der Name, unter dem das Modell gespeichert werden soll.

**Tabellenname.** Der Name der Datenbanktabelle, in der das Modell gespeichert werden soll, laut Angabe im Modellierungsknoten. Der Name wird nur zu Referenzzwecken angegeben und kann hier nicht geändert werden.

**Eingabefelder einschließen.** Wenn diese Option ausgewählt ist (Standardvorgabe) werden alle ursprünglichen Eingabefelder nach unten weitergegeben, wobei das zusätzliche Modellierungsfeld (bzw. die zusätzlichen Modellierungsfelder) an die einzelnen Datenzeilen angehängt werden. Wenn Sie dieses Kontrollkästchen deaktivieren, werden nur das Feld "Datensatz-ID" und die zusätzlichen Modellierungsfelder weitergeleitet, sodass der Stream schneller abläuft.

**Berechnen Sie Wahrscheinlichkeiten zugewiesener Klassen für Bewertungsdatensätze.** (Nur Decision Tree) Wenn diese Option ausgewählt ist (Standardeinstellung) enthalten die zusätzlichen Modellierungsfelder neben dem Vorhersagefeld auch eine Konfidenzfeld (also ein Wahrscheinlichkeitsfeld). Wenn Sie dieses Kontrollkästchen deaktivieren wird nur das Vorhersagefeld erstellt.

### **Netezza-Modell-Nugget vom Typ "K-Means"**

Modell-Nuggets für K-Means-Modelle enthalten alle Informationen, die vom Cluster-Modell erfasst wurden, sowie Informationen zu den Trainingsdaten und dem Schätzvorgang.

Wenn Sie einen Stream ausführen, der ein Modell-Nugget vom Typ "K-Means" enthält, fügt dieser Knoten zwei neue Felder hinzu, die die Cluster-Mitgliedschaft und die Entfernung vom zugewiesenen Cluster-Zentrum für den betreffenden Datensatz enthalten. Die neuen Feldnamen werden durch Präfigierung von *\$KM-* für die Cluster-Mitgliedschaft und *\$KMD-* für die Entfernung vom Cluster-Zentrum aus dem Modellnamen abgeleitet. Beispiel: Wenn das Modell den Namen *Kmeans* trägt, erhalten die neuen Felder die Namen *\$KM-Kmeans* und *\$KMD-Kmeans*.

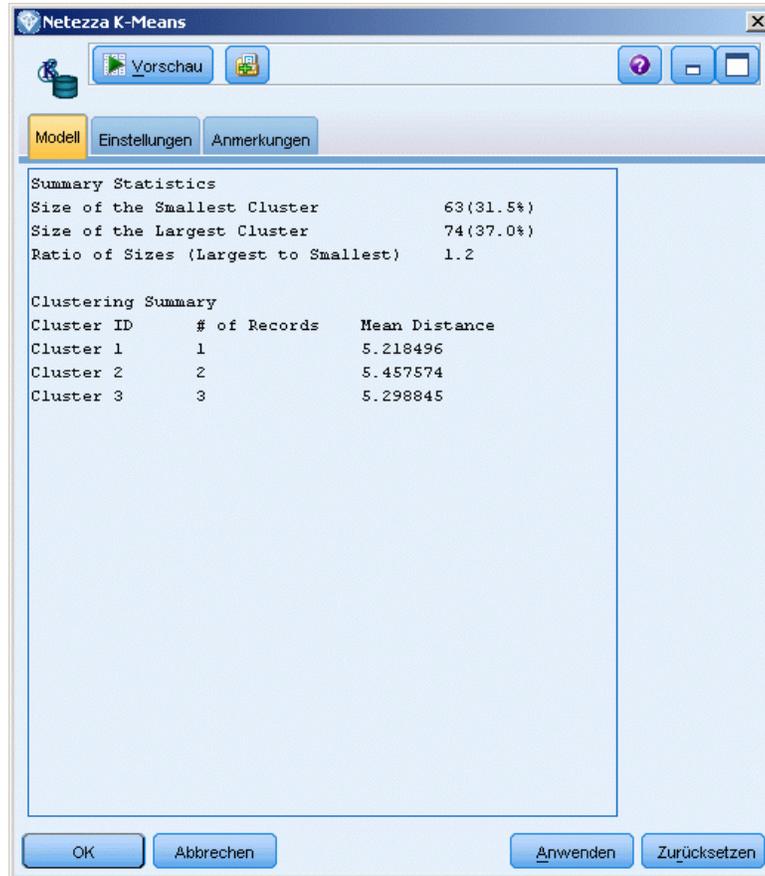
So können Sie die vollständige Modellausgabe für das Scoring anzeigen:

- ▶ Verbinden Sie einen Tabellenknoten mit dem Modell-Nugget.
- ▶ Öffnen Sie den Tabellenknoten.
- ▶ Klicken Sie auf Ausführen.

Blättern Sie im Tabellenausgabefenster nach rechts, um die Vorhersagen anzuzeigen.

### Netezza-K-Means-Nugget - Registerkarte "Modell"

Abbildung 6-12  
K-Means-Modellausgabe



Die Modellausgabe wird wie folgt auf der Registerkarte "Modell" angezeigt.

**Übersichtsstatistiken.** Für den kleinsten und den größten Cluster werden die Anzahl an Datensätzen und der Prozentsatz des Daten-Sets angezeigt, der auf diese Cluster entfällt. In der Liste wird auch das Größenverhältnis zwischen dem größten und dem kleinsten Cluster angegeben.

**Clusterübersicht.** Listet die vom Algorithmus erstellten Cluster auf. Für jeden Cluster wird in der Tabelle die Anzahl der Datensätze in diesem Cluster angezeigt, sowie die mittlere Entfernung vom Clusterzentrum für diese Datensätze.

### Netezza-K-Means-Nugget - Registerkarte "Einstellungen"

Auf der Registerkarte "Einstellungen" können Sie einige Optionen für das Scoring des Modells festlegen.

Abbildung 6-13  
Einstellungen für K-Means-Modelle

**Allgemein.** Dieser Abschnitt enthält einige allgemeine Optionen für das Scoring von Modellen.

- **Modellname.** Der Name, unter dem das Modell gespeichert werden soll.
- **Tabellenname.** Der Name der Datenbanktabelle, in der das Modell gespeichert werden soll, laut Angabe im Modellierungsknoten. Der Name wird nur zu Referenzzwecken angegeben und kann hier nicht geändert werden.

**Netezza-DB-Server-Details.** Hier geben Sie die Verbindungsdetails für die für das Modell zu verwendende Datenbank an.

- **Oberhalb gelegene Verbindung verwenden.** (Standardeinstellung) Verwendet die Verbindungsdetails, die in einem oberhalb gelegenen Knoten, beispielsweise dem Datenbank-Quellenknoten, angegeben sind. *Hinweis:* Diese Option funktioniert nur, wenn alle oberhalb gelegenen Knoten SQL-Pushback verwenden können. In diesem Fall müssen die Daten nicht aus der Datenbank verschoben werden, da die SQL alle oberhalb gelegenen Knoten vollständig implementiert.
- **Daten in Verbindung verschieben.** Dient zum Verschieben der Daten in die hier angegebene Datenbank. Dadurch kann die Modellierung funktionieren, wenn sich die Daten in einer anderen IBM Netezza-Datenbank, einer Datenbank eines anderen Anbieters oder in einer Textdatei befinden. Darüber hinaus werden die Daten in die hier angegebene Datenbank zurückverschoben, wenn die Daten extrahiert wurden, da ein Knoten kein SQL-Pushback durchgeführt hat. Klicken Sie auf die Schaltfläche *Bearbeiten*, um eine Verbindung zu suchen und auszuwählen. *Vorsicht:* IBM® Netezza® Analytics wird in der Regel mit sehr großen Daten-Sets verwendet. Das Übertragen großer Datenmengen zwischen Datenbanken bzw. aus einer Datenbank und wieder zurück kann sehr zeitaufwendig sein und sollte nach Möglichkeit vermieden werden.

**Für Scoring bereitstellen.** Dieser Abschnitt enthält einige spezifische Optionen für das Scoring von Modellen.

- **Eingabefelder einschließen.** Wenn diese Option ausgewählt ist (Standardvorgabe) werden alle ursprünglichen Eingabefelder nach unten weitergegeben, wobei das zusätzliche Modellierungsfeld (bzw. die zusätzlichen Modellierungsfelder) an die einzelnen Datenzeilen angehängt werden. Wenn Sie dieses Kontrollkästchen deaktivieren, werden nur das Feld "Datensatz-ID" und die zusätzlichen Modellierungsfelder weitergeleitet, sodass der Stream schneller abläuft.
- **Distanzmaß.** Die zur Messung des Abstands zwischen Datenpunkten zu verwendende Methode. Größere Abstände deuten auf größere Unähnlichkeiten hin. Folgende Optionen stehen zur Auswahl:
  - Euklidisch.** (Standardvorgabe). Der Abstand zwischen zwei Punkten wird anhand einer geradlinigen Verbindung zwischen den Punkten berechnet.
  - Manhattan.** Der Abstand zwischen zwei Punkten wird als Summe der absoluten Unterschiede zwischen ihren Koordinaten berechnet.
  - Canberra.** Ähnlich wie die Manhattan-Distanz, jedoch empfindlicher für Datenpunkte, die näher am Ursprung liegen.

**Maximum.** Der Abstand zwischen zwei Punkten wird als der größte Unterschied in einer beliebigen Koordinatendimension berechnet.

## **Hinweise**

This information was developed for products and services offered worldwide.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

*IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785, U.S.A.*

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

*Intellectual Property Licensing, Legal and Intellectual Property Law, IBM Japan Ltd., 1623-14, Shimotsuruma, Yamato-shi, Kanagawa 242-8502 Japan.*

**Der folgende Absatz gilt nicht für Großbritannien oder andere Länder, in denen derartige Bestimmungen nicht mit dem dort geltenden Recht vereinbar sind.** SPSS INC., AN IBM COMPANY, ÜBERNIMMT FÜR DIE VORLIEGENDE DOKUMENTATION KEINERLEI GEWÄHRLEISTUNG IRGENDWELCHER ART, WEDER AUSDRÜCKLICH NOCH STILLSCHWEIGEND, EINSCHLIESSLICH (JEDOCH NICHT DARAUF BEGRENZT) DER STILLSCHWEIGENDEN GEWÄHRLEISTUNGEN IN BEZUG AUF DIE NICHTVERLETZUNG VON RECHTEN DRITTER, AUF HANDELSÜBLICHKEIT ODER DIE EIGNUNG FÜR EINEN BESTIMMTEN ZWECK. Einige Staaten lassen bei bestimmten Transaktionen keine Ausschlussklauseln ausdrücklicher oder stillschweigender Gewährleistungen zu, sodass diese Erklärung möglicherweise nicht auf Sie zutrifft.

Diese Informationen können technische Ungenauigkeiten oder typografische Fehler enthalten. An den hierin enthaltenen Informationen werden in regelmäßigen Abständen Änderungen vorgenommen, die in spätere Ausgaben der Publikation eingearbeitet werden. SPSS Inc. kann jederzeit ohne Vorankündigung Verbesserungen und/oder Veränderungen an den in dieser Publikation beschriebenen Produkten und/oder Programmen vornehmen.

Alle in diesen Ausführungen enthaltenen Verweise auf Websites, die nicht zu SPSS bzw. IBM gehören, dienen lediglich der Information. Die Nennung bedeutet nicht, dass SPSS bzw. IBM den Inhalt dieser Websites unterstützen. Das Material auf diesen Websites ist kein Bestandteil des Materials für dieses SPSS Inc.-Produkt. Sie verwenden diese Websites auf eigenes Risiko.

Wenn Sie Informationen an IBM bzw. SPSS senden, räumen Sie IBM und SPSS das nicht ausschließliche Recht ein, die Informationen in jeglicher Form zu verwenden bzw. weiterzugeben, die dem Unternehmen geeignet erscheint, ohne dass ihm daraus Verbindlichkeiten Ihnen gegenüber entstehen.

Informationen zu Nicht-SPSS-Produkten stammen von den Herstellern dieser Produkte, ihren veröffentlichten Verlautbarungen oder aus anderen öffentlich verfügbaren Quellen. SPSS hat diese Produkte nicht getestet und kann daher die Richtigkeit der Angaben zu Leistung und Kompatibilität oder anderer Behauptungen in Bezug auf Nicht-SPSS-Produkte nicht bestätigen. Fragen zu den Fähigkeiten von Nicht-SPSS-Produkten sind an die Hersteller dieser Produkte zu richten.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact:

*IBM Software Group, Attention: Licensing, 233 S. Wacker Dr., Chicago, IL 60606, USA.*

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The licensed program described in this document and all licensed material available for it are provided by IBM under terms of the IBM Customer Agreement, IBM International Program License Agreement or any equivalent agreement between us.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

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

All statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

Diese Informationen enthalten Beispiele für Daten und Berichte, die in alltäglichen Betriebsabläufen verwendet werden. Um sie möglichst umfassend darzulegen, enthalten die Beispiele Namen von Einzelpersonen, Unternehmen, Marken und Produkten. Alle diese Namen sind frei erfunden und jegliche Ähnlichkeit mit Namen und Adressen, die von einem tatsächlichen Handelsunternehmen verwendet werden, ist rein zufällig.

If you are viewing this information softcopy, the photographs and color illustrations may not appear.

**Marken**

IBM, das IBM-Logo und ibm.com sind Marken von IBM Corporation, die in vielen Ländern weltweit eingetragen sind. Eine aktuelle Liste der IBM-Marken finden Sie im Internet unter <http://www.ibm.com/legal/copytrade.shtml>.

SPSS ist eine Marke von SPSS Inc., an IBM Company, die in vielen Ländern weltweit eingetragen sind.

Adobe, das Adobe-Logo, PostScript und das PostScript-Logo sind entweder registrierte Marken oder Marken von Adobe Systems Incorporated in den USA und/oder anderen Ländern.

IT Infrastructure Library ist eine eingetragene Marke der Central Computer and Telecommunications Agency, die nun zum Office of Government Commerce gehört.

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 und 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 von Microsoft Corporation in den USA und/oder anderen Ländern.

ITIL ist eine eingetragene Marke und eine eingetragene Gemeinschaftsmarke des Office of Government Commerce und ist beim U.S. Patent and Trademark Office eingetragen.

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

Cell Broadband Engine ist eine Marke von Sony Computer Entertainment, Inc. in den USA und/oder anderen Ländern und wird im Rahmen einer Lizenz dieses Unternehmens verwendet.

Java und alle Java-basierten Marken und Logos sind Marken von Sun Microsystems, Inc. in den USA und/oder anderen Ländern.

Linear Tape-Open, LTO, the LTO Logo, Ultrium, and the Ultrium logo are trademarks of HP, IBM Corp. and Quantum in the U.S. and other countries.

Weitere Produkt- oder Servicenamen können Marken von IBM, SPSS oder anderen Unternehmen sein.

- A Priori
  - Microsoft, 28
  - Oracle Data Mining, 83, 86
- A-priori-Wahrscheinlichkeit
  - Oracle Data Mining, 69
- Adaptive Bayes Network
  - Oracle Data Mining, 63–65
- Analysis Services
  - Beispiele, 45
  - Decision Trees (Entscheidungsbäume), 45
  - Integrieren mit IBM SPSS Modeler, 6, 13
  - Verwalten von Modellen, 19
- Anwendungsbeispiele, 3
- Anzahl der Cluster
  - Oracle k-Means, 79
  - Oracle O-Cluster, 77
- Assoziationsmodellierung
  - InfoSphere Warehouse Data Mining, 125
- Assoziationsregelmodelle
  - Microsoft, 28
- Assoziationsregeln
  - Expertenoptionen, 29
  - Modelloptionen, 21
  - Scoring – Serveroptionen, 37
  - Scoring - Übersichtsoptionen, 39
  - Serveroptionen, 21
- Attribute Importance (AI)
  - Oracle Data Mining, 89–91
  
- Beispiele
  - Anwendungshandbuch, 3
  - Database-Mining, 44–48, 51, 100, 154–157, 159
  - Übersicht, 4
- Bereitstellung, 51, 105, 159
- Bewertung, 9
  
- Clustering
  - Expertenoptionen, 24
  - InfoSphere Warehouse Data Mining, 139
  - Modelloptionen, 21
  - Scoring – Serveroptionen, 37
  - Scoring - Übersichtsoptionen, 39
  - Serveroptionen, 21
- Clustering-Knoten
  - InfoSphere Warehouse Data Mining, 139
  
- Data Audit-Knoten, 46, 100, 155
- Database-Mining
  - Beispiel, 44, 154
  - Datenvorbereitung, 8
  - Erstellen von Modellen, 8
  - Konfiguration, 15
  - Optimierungsoptionen, 8
  - Verwendung von IBM SPSS Modeler, 7
- Datei *tnsnames.ora*, 55
  
- Datenbank-Modellierung
  - Analysis Services, 6
  - IBM InfoSphere Warehouse (ISW), 6
  - IBM Netezza Analytics, 161–162, 164
  - Oracle, 53–54, 57, 59
  - Oracle Data Miner, 6
- DB2
  - Verwalten von Modellen, 117
- Distanzfunktion
  - Oracle k-Means, 79
- Dokumentation, 3
- DSN
  - Konfigurieren, 15
  
- eindeutiges Feld
  - Oracle Adaptive Bayes Network, 64
  - Oracle Apriori, 75, 86
  - Oracle Data Mining, 58
  - Oracle k-Means, 79
  - Oracle MDL, 89
  - Oracle Naive Bayes, 62
  - Oracle NMF, 81
  - Oracle O-Cluster, 77
  - Oracle Support Vector Machine, 66
- Einzelfunktionsmodelle
  - Oracle Adaptive Bayes Network, 64
- Entropie-Unreinheitsmaß, 171
- Entscheidungsbaum
  - IBM Netezza Analytics, 167–168, 170, 172, 174, 178–179
  - Oracle Data Mining, 74–76
- Entscheidungsbaum-Modelle
  - InfoSphere Warehouse Data Mining, 123
- Entscheidungsbäume
  - Expertenoptionen, 23
  - Microsoft Analysis Services, 12, 15, 36
  - Modelloptionen, 21
  - Scoring – Serveroptionen, 37
  - Scoring - Übersichtsoptionen, 39
  - Serveroptionen, 21
- epsilon
  - Oracle Support Vector Machine, 68
- Erstellungsoptionen
  - IBM Netezza Analytics, 170, 172, 174, 177
- evaluation, 48, 102, 157
- Exploration, 46, 100, 155
- export
  - Analysis Services-Modelle, 44
  - DB2-Modelle, 119
  
- Fehlklassifizierungskosten
  - Entscheidungsbäume, 60, 122
  - Oracle, 60
- Feldoptionen
  - IBM Netezza Analytics, 168, 175

- Gauss'scher Kernel
  - Oracle Support Vector Machine, 66
- Gini-Unreinheitsmaß, 171
- Hostname
  - Oracle-Verbindung, 55
- IBM
  - Assoziationsmodellierung, 106
  - Demografischer Cluster, Modellierung, 106
  - Entscheidungsbaum-Modellierung, 106
  - Kohonen-Cluster-Modellierung, 106
  - Lineare Regression, Modellierung, 106
  - Logistische Regression, Modellierung, 106
  - Naive Bayes-Modellierung, 106
  - Polynomiale Regression, Modellierung, 106
  - Regressionsmodellierung, 106
  - Sequenzmodellierung, 106
  - Verwalten von Modellen, 117
  - Zeitreihenmodellierung, 106
- IBM InfoSphere Warehouse (ISW)
  - Integrieren mit IBM SPSS Modeler, 6
- IBM Netezza Analytics
  - Decision Trees (Entscheidungsbäume), 167
  - Entscheidungsbaum-Erstellungsoptionen, 170, 172, 174
  - Entscheidungsbaummodell-Nugget, 178–179
  - K-Means, 175
  - K-Means-Erstellungsoptionen, 177
  - K-Means-Feldoptionen, 175
  - K-Means-Modell-Nugget, 181–182
  - Konfigurieren mit IBM SPSS Modeler, 161–162, 164
  - Modelloptionen, 166
  - Optionen für Entscheidungsbaumfelder, 168
  - Verwalten von Modellen, 178
- IBM SPSS Modeler, 1
  - Database-Mining, 7
  - Dokumentation, 3
- IBM SPSS Modeler Solution Publisher
  - Oracle Data Mining-Modelle, 58
- IBM SPSS Text Analytics, 2
- InfoSphere Warehouse (IBM), siehe ISW, 107
- InfoSphere Warehouse Data Mining
  - Assoziationsmodellierung, 125
  - Beispiel-Streams, 154
  - Entscheidungsbäume, 123
  - Modell-Nuggets, 151
  - Regressionsknoten, 134
  - Sequenzknoten, 131
  - Taxonomie, 128
- ISW
  - Integrieren mit IBM SPSS Modeler, 107
  - ODBC-Verbindung, 107
  - Server (Registerkarte), 119
- k-Means
  - IBM Netezza Analytics, 175, 177
  - Oracle Data Mining, 78–80
- K-Means
  - IBM Netezza Analytics, 181–182
- Kategorieneditor
  - ISW-Assoziationsknoten, 130
- Klassieren der Daten
  - Oracle-Modelle, 98
- Knoten
  - erzeugen, 44
  - Knoten erzeugen, 44
- Komplexitätsfaktor
  - Oracle Support Vector Machine, 68
- Komplexitätsstrafe, 23–29, 32
- Konvergenztoleranz
  - Oracle Support Vector Machine, 68
- Kosten
  - Oracle, 60
- Kreuzvalidierung
  - Oracle Naive Bayes, 61
- Lineare Regression
  - Expertenoptionen, 26
  - Modelloptionen, 21
  - Scoring – Serveroptionen, 37
  - Scoring - Übersichtsoptionen, 39
  - Serveroptionen, 21
- Linearer Kernel
  - Oracle Support Vector Machine, 66
- Logistische Regression
  - Expertenoptionen, 28
  - Modelloptionen, 21
  - Scoring – Serveroptionen, 37
  - Scoring - Übersichtsoptionen, 39
  - Serveroptionen, 21
- Logistischer Regressionsknoten
  - InfoSphere Warehouse Data Mining, 145
- Marken, 187
- MDL, 63
- Microsoft
  - Analysis Services, 12, 15, 36
  - Assoziationsregelmodellierung, 12, 15, 36
  - Cluster-Modellierung, 12, 15, 36
  - Entscheidungsbaum-Modellierung, 12, 15, 36
  - Lineare Regression, 12
  - Lineare Regression, Modellierung, 15, 36
  - Logistische Regression, 12
  - Logistische Regression, Modellierung, 15, 36
  - Naive Bayes-Modellierung, 12, 15, 36
  - Neuronale Netzwerke, Modellierung, 15, 36
  - Neuronales Netzwerk, 12
  - Sequenz-Clustering, 12
  - Verwalten von Modellen, 19
- Microsoft Analysis Services, 40, 43–44
  - Integrieren mit IBM SPSS Modeler, 13
- Microsoft SQL Server
  - Integrieren mit IBM SPSS Modeler, 13

- Min-Max
  - Normalisieren von Daten, 67, 98
- Minimum Description Length, 63
- Minimum Description Length (MDL)
  - Oracle Data Mining, 87–88
- Modell-Nuggets
  - IBM Netezza Analytics, 178–179, 181–182
  - InfoSphere Warehouse Data Mining, 151
- Modell-Scoring
  - InfoSphere Warehouse Data Mining, 115
- Modelle
  - Analysis Services verwalten, 19
  - Auflistung von DB2, 117
  - Durchsuchen von DB2, 118
  - Durchsuchen von Oracle, 64
  - evaluation, 48, 102, 157
  - exportieren, 10
  - in der Datenbank erstellen, 8
  - innerhalb der Datenbank scoren, 9
  - Konsistenzprobleme, 11
  - speichern, 10
  - Verwalten von DB2, 117
- Modellierung innerhalb der Datenbank, 39
  - Analysis Services, 6
  - IBM InfoSphere Warehouse (ISW), 6
  - Oracle Data Miner, 6
- Modellierungsknoten
  - Datenbankinterne Modellbildung für ISW, 107
  - Microsoft Decision Trees, 19
  - Microsoft Naive Bayes, 19
  - Microsoft Time Series, 19
  - Microsoft, lineare Regression, 19
  - Microsoft, logistische Regression, 19
  - Microsoft, neuronales Netzwerk, 19
  - Microsoft-Assoziationsregeln, 19
  - Microsoft-Clusterbildung, 19
  - Microsoft-Sequenz-Clustering, 19
  - Modellierung innerhalb der Datenbank, 9, 12, 15, 19, 36
- Modelloptionen
  - IBM Netezza Analytics, 166
- Multifunktionsmodelle
  - Oracle Adaptive Bayes Network, 64
  
- Naive Bayes
  - Expertenoptionen, 25
  - InfoSphere Warehouse Data Mining, 144
  - Modelloptionen, 21
  - Oracle Data Mining, 61–62
  - Scoring – Serveroptionen, 37
  - Scoring - Übersichtsoptionen, 39
  - Serveroptionen, 21
- Naive Bayes-Modelle
  - Oracle Adaptive Bayes Network, 64
- Netezza Analytics, 161
- Neuronales Netzwerk
  - Expertenoptionen, 27
  - Modelloptionen, 21
- Scoring – Serveroptionen, 37
- Scoring - Übersichtsoptionen, 39
- Serveroptionen, 21
- NMF
  - Oracle Data Mining, 80–82
- Normalisieren von Daten
  - Oracle-Modelle, 98
- Normalisierungsmethode
  - Oracle k-Means, 79
  - Oracle NMF, 81
  - Oracle Support Vector Machine, 67
  
- O-Cluster
  - Oracle Data Mining, 77–78
- ODBC
  - ISW konfigurieren, 107
  - Konfigurieren, 15
  - Konfigurieren für IBM Netezza Analytics, 161–162, 164
  - Konfigurieren für Oracle, 53–54, 57, 59
  - SQL Server konfigurieren, 16
- ODM. *Siehe* Oracle Data Mining, 73
- Optimierung
  - SQL-Erzeugung, 6
- Oracle Data Miner, 96
  - Integrieren mit IBM SPSS Modeler, 6
- Oracle Data Mining, 53
  - A Priori, 83, 86
  - Adaptive Bayes Network, 63–65
  - Attribute Importance (AI), 89–91
  - Beispiele, 98–102, 105
  - Daten vorbereiten, 98
  - Entscheidungsbaum, 74–76
  - Fehlklassifizierungskosten, 94
  - k-Means, 78–80
  - Konfigurieren mit IBM SPSS Modeler, 53–54, 57, 59
  - Konsistenzprüfung, 92
  - Minimum Description Length (MDL), 87–88
  - Naive Bayes, 61–62
  - NMF, 80–82
  - O-Cluster, 77–78
  - Support Vector Machine, 66, 68
  - Verallgemeinerte lineare Modelle (GLM), 70–73
  - Verwalten von Modellen, 92–94
  
- Pairwise-Grenzwert
  - Oracle Naive Bayes, 63
- Partitionen
  - Modellerstellung, 31, 62, 64, 90, 127, 132, 136, 141, 145–146
- Partitionieren von Daten, 85
- Partitionsfelder
  - auswählen, 85
- Port
  - Oracle-Verbindung, 55
- Power-Optionen
  - ISW Data Mining, 121

- 
- Publisher-Knoten
    - Oracle Data Mining-Modelle, 58
  
  - Rechtliche Hinweise, 185
  - Reduzierte Naive Bayes-Modelle
    - Oracle Adaptive Bayes Network, 64
  - Regressionsknoten
    - InfoSphere Warehouse Data Mining, 134
  
  - Schlüssel
    - Modellschlüssel, 11
  - Sequenz-Clustering
    - Modelloptionen, 21
  - Sequenz-Clustering (Microsoft), 34
    - Expertenoptionen, 36
    - Feldoptionen, 34
  - Sequenzknoten
    - InfoSphere Warehouse Data Mining, 131
  - Server
    - Ausführen von Analysis Services, 21, 37, 39
  - Server (Registerkarte)
    - ISW, 119
  - SID
    - Oracle-Verbindung, 55
  - Singleton-Grenzwert
    - Oracle Naive Bayes, 63
  - Solution Publisher
    - Oracle Data Mining-Modelle, 58
  - Split-Kriterium
    - Oracle k-Means, 79
  - SPSS Modeler Server, 1
  - SQL Server, 21, 37, 39
    - Integrieren mit IBM SPSS Modeler, 13
    - Konfigurieren, 15
    - ODBC-Verbindung, 16
  - SQL-Erzeugung, 6, 8
  - SQL-Optimierung. *Siehe* SQL-Erzeugung, 6
  - SQL-Pushback. *Siehe* SQL-Erzeugung, 6
  - Standardabweichung
    - Oracle Support Vector Machine, 68
  - Streams
    - Beispiele für InfoSphere Warehouse Data Mining, 154
  - Stufen der
    - Datenbankinterne Modellbildung für ISW, 107
    - Modellierung innerhalb der Datenbank, 9, 12, 15, 19, 36
  - Support Vector Machine
    - Oracle Data Mining, 66, 68
  - SVM. *Siehe* Support Vector Machine, 66
  
  - Taxonomie
    - InfoSphere Warehouse Data Mining, 128
  
  - Unreinheitsmaße
    - Netezza-Entscheidungsbaum, 171
  - Unreinheitsmetrik
    - Oracle Apriori, 75
  
  - Verallgemeinerte lineare Modelle (GLM)
    - Oracle Data Mining, 70–73
  
  - Z-Werte
    - Normalisieren von Daten, 67, 98
  - Zeitreihen
    - InfoSphere Warehouse Data Mining, 146–150
  - Zeitreihen (Microsoft), 30
    - Einstellungsoptionen, 33
    - Expertenoptionen, 32
    - Modelloptionen, 31