

*IBM SPSS Modeler 18.0 —
podręcznik eksploracji w bazie
danych*

IBM

Uwaga

Przed skorzystaniem z niniejszych informacji oraz produktu, którego one dotyczą, należy zapoznać się z informacjami zamieszczonymi w sekcji “Uwagi” na stronie 127.

Informacje o produkcie

Niniejsze wydanie publikacji dotyczy wersji 18, wydania 0, modyfikacji 0 produktu IBM SPSS Modeler oraz wszystkich następnych wersji i modyfikacji do czasu, aż w kolejnym wydaniu publikacji zostanie zawarta informacja o stosownej zmianie.

Spis treści

Przedmowa	vii
----------------------------	------------

Rozdział 1. O programie IBM SPSS

Modeler	1
--------------------------	----------

Produkty IBM SPSS Modeler	1
IBM SPSS Modeler	1
IBM SPSS Modeler Server	1
IBM SPSS Modeler Administration Console	2
IBM SPSS Modeler Batch	2
IBM SPSS Modeler Solution Publisher	2
IBM SPSS Modeler Server Adapters for IBM SPSS Collaboration and Deployment Services	2
Wydania programu IBM SPSS Modeler	2
Dokumentacja IBM SPSS Modeler	3
Dokumentacja SPSS Modeler Professional	3
Dokumentacja SPSS Modeler Premium	4
Przykłady aplikacji	4
Folder Demos	5
Monitorowanie wykorzystania licencji	5

Rozdział 2. Eksploracja w bazie danych 7

Przegląd informacji o modelowaniu w bazie danych	7
Co jest potrzebne	7
Budowanie modelu	8
Przygotowanie danych	8
Ocenianie modelu	8
Eksportowanie i zapisywanie modeli bazodanowych	9
Spójność modelu	9
Wyświetlanie i eksportowanie wygenerowanego kodu SQL	9

Rozdział 3. Modelowanie w bazie danych przy użyciu programu Microsoft Analysis Services 11

Produkt IBM SPSS Modeler i usługi Microsoft Analysis Services	11
Wymagania dotyczące integracji z programem Microsoft Analysis Services	12
Aktywacja integracji z usługami Analysis Services	13
Budowanie modeli za pomocą usług Analysis Services	15
Zarządzanie modelami usług Analysis Services	15
Ustawienia wspólne dla wszystkich węzłów algorytmów	17
Opcje zaawansowane MS Decision Tree	17
Opcje zaawansowane MS Clustering	17
Opcje zaawansowane MS Naive Bayes	18
Opcje zaawansowane MS Linear Regression	18
Opcje zaawansowane MS Neural Network	18
Opcje zaawansowane MS Logistic Regression	18
Węzeł reguł asocjacyjnych MS	18
Węzeł MS Time Series	19
Węzeł MS Sequence Clustering	20
Ocenianie modelami usług Analysis Services	21

Ustawienia wspólne dla wszystkich algorytmów usług Analysis Services	21
Model użytkowy MS Time Series	22
Model użytkowy MS Sequence Clustering	23
Eksportowanie modeli i generowanie węzłów	24
Przykłady eksploracji w usługach Analysis Services	24
Przykładowe strumienie: Drzewa decyzyjne	24

Rozdział 4. Modelowanie w bazie danych z użyciem rozwiązania Oracle Data Mining 27

Informacje o rozwiązaniu Oracle Data Mining	27
Wymagania dotyczące integracji z systemem Oracle	27
Aktywacja integracji z produktem Oracle	28
Budowanie modeli z użyciem rozwiązania Oracle Data Mining	29
Modele Oracle — opcje serwera	30
Koszty błędnej klasyfikacji	30
Oracle Naive Bayes	31
Opcje modelu Naive Bayes	31
Opcje zaawansowane Naive Bayes	31
Oracle Adaptive Bayes	32
Opcje modelu Adaptive Bayes	32
Opcje zaawansowane Adaptive Bayes	33
Oracle Support Vector Machine (SVM)	33
Opcje modelu Oracle SVM	33
Opcje zaawansowane Oracle SVM	34
Opcje wag dla Oracle SVM	35
Uogólnione modele liniowe Oracle (GLM)	35
Opcje modelu Oracle GLM	35
Opcje zaawansowane Oracle GLM	36
Opcje wag dla Oracle GLM	36
Oracle Decision Tree	37
Opcje modelu drzewa decyzyjnego	37
Opcje zaawansowane drzewa decyzyjnego	38
Oracle O-Cluster	38
Opcje modelu O-Cluster	38
Opcje zaawansowane O-Cluster	39
Oracle K-średnie	39
Opcje modelu K-średnie	39
Opcje zaawansowane K-średnich	39
Oracle Nonnegative Matrix Factorization (NMF)	40
Opcje modelu NMF	40
Opcje zaawansowane NMF	40
Oracle Apriori	41
Opcje zmiennych Apriori	41
Opcje modelu Apriori	42
Oracle Minimum Description Length (MDL)	42
Opcje modelu MDL	43
Oracle Attribute Importance (AI)	43
Opcje modelu AI	43
Opcje doboru AI	44
Model użytkowy AI — karta Model	44
Zarządzanie modelami Oracle	44
Karta serwera modelu użytkowego Oracle	44

Karta podsumowania modelu użytkowego Oracle	45
Karta ustawień modelu użytkowego Oracle	45
Lista modeli Oracle	45
Oracle Data Miner.	46
Przygotowywanie danych	46
Przykłady dotyczące Oracle Data Mining	47
Przykładowy strumień: Wczytywanie danych.	47
Przykładowy strumień: Eksploracja danych	48
Przykładowy strumień: Budowa modelu	48
Przykładowy strumień: Ocena modelu	48
Przykładowy strumień: Wdrożenie modelu	48
Rozdział 5. Modelowanie w bazie za pomocą produktu IBM InfoSphere Warehouse	49
IBM InfoSphere Warehouse i produkt IBM SPSS Modeler	49
Wymagania dotyczące integracji z produktem IBM InfoSphere Warehouse	49
Włączanie integracji z produktem IBM InfoSphere Warehouse	50
Budowanie modelu za pomocą produktu IBM InfoSphere Warehouse Data Mining	53
Ocenianie i wdrażanie modelu	53
Zarządzanie modelami DB2	54
Lista modeli bazy danych	54
Przeglądanie modeli	55
Eksportowanie modeli i generowanie węzłów	55
Ustawienia węzłów wspólne dla wszystkich algorytmów.	55
ISW Decision Tree	57
Opcje modelu ISW Decision Tree	57
Opcje zaawansowane ISW Decision Tree	58
ISW Association	58
Opcje zmiennej ISW Association	58
Opcje modelu asocjacyjnego ISW	60
Opcje zaawansowane ISW Association	60
Opcje systematyki ISW	60
ISW Sequence	61
Opcje modelu ISW Sequence	62
Opcje zaawansowane ISW Sequence	62
ISW Regression	63
Opcje modelu ISW Regression	64
Opcje zaawansowane ISW Regression	64
ISW Clustering.	65
Opcje modelu ISW Clustering	66
Opcje zaawansowane ISW Clustering	67
ISW Naive Bayes	68
Opcje modelu ISW Naive Bayes	68
ISW Logistic Regression	68
Opcje modelu ISW Logistic Regression	68
ISW Time Series	68
Opcje zmiennych ISW Time Series.	69
Opcje modelu ISW Time Series.	69
Opcje zaawansowane ISW Time Series	70
Wyświetlanie modeli ISW Time Series	70
Modele użytkowe ISW Data Mining	70
Karta serwera modelu użytkowego ISW	70
Karta ustawień modelu użytkowego ISW	71
Karta podsumowania modelu użytkowego ISW	71
Przykłady dotyczące ISW Data Mining	71

Przykładowy strumień: Wczytywanie danych.	72
Przykładowy strumień: Eksploracja danych	72
Przykładowy strumień: Budowa modelu	72
Przykładowy strumień: Ocena modelu	72
Przykładowy strumień: Wdrożenie modelu	73

Rozdział 6. Modelowanie w bazie za pomocą produktu IBM Netezza Analytics

IBM SPSS Modeler i IBM Netezza Analytics.	75
Wymagania dotyczące integracji z produktem IBM Netezza Analytics	75
Aktywacja integracji z produktem IBM Netezza Analytics	76
Konfigurowanie produktu IBM Netezza Analytics	76
Tworzenie źródła ODBC dla produktu IBM Netezza Analytics	76
Aktywacja integracji z produktem IBM Netezza Analytics w produkcie IBM SPSS Modeler	77
Włączenie opcji generowania i optymalizacji kodu SQL	78
Budowanie modeli za pomocą produktu IBM Netezza Analytics	78
Modele Netezza — opcje zmiennych	79
Modele Netezza — opcje serwera	79
Modele Netezza — opcje modelu	80
Zarządzanie modelami Netezza	80
Lista modeli bazy danych	80
Netezza Regression Tree	81
Opcje budowania drzewa regresji Netezza — wzrost drzewa	81
Opcje budowania drzewa regresji Netezza — przycinanie drzewa	82
Netezza Divisive Clustering	82
Opcje zmiennej grupowania dzielącego Netezza	83
Opcje budowania grupowania dzielącego Netezza	83
Netezza Generalized Linear	84
Opcje zmiennej modelowania z użyciem uogólnionego modelu liniowego Netezza	84
Opcje zmiennej modelowania z użyciem uogólnionego modelu liniowego Netezza — informacje ogólne.	85
Opcje zmiennej modelowania z użyciem uogólnionego modelu liniowego Netezza — interakcja	85
Opcje uogólnionego modelu liniowego Netezza — ocje oceniania.	86
Drzewa decyzyjne Netezza	87
Wagi instancji i wagi klas	87
Opcje zmiennej drzewa decyzyjnego Netezza.	88
Opcje budowania drzewa decyzyjnego Netezza	88
Netezza Linear Regression	90
Opcje budowania regresji liniowej Netezza	90
Netezza KNN	90
Opcje modelu KNN Netezza — informacje ogólne	90
Opcje modelu KNN Netezza — opcje oceny	91
Netezza K-Means	92
Opcje zmiennej K-średnie Netezza	92
Karta Opcje budowania K-średnie Netezza	92
Netezza Naive Bayes	93
Netezza Bayes Net	93
Opcje zmiennej sieci Bayesa Netezza	93
Opcje budowania zmiennej Sieci Bayesa Netezza	94
Netezza Time Series	94

Interpolacja wartości w szeregu czasowym Netezza	95
Opcje zmiennych szeregu czasowego Netezza	96
Opcje budowania szeregu czasowego Netezza	97
Opcje modelu szeregów czasowych Netezza	99
Netezza TwoStep	99
Opcje zmiennej Dwustopniowa w Netezza	100
Opcje budowania zmiennej Dwustopniowa w Netezza	100
Netezza PCA	101
Opcje zmiennej PCA Netezza	101
Opcje budowania PCA Netezza	101
Zarządzanie modelami IBM Netezza Analytics	102
Ocenianie modeli IBM Netezza Analytics	102
Karta Serwer modelu użytkowego Netezza	102
Modele użytkowe drzewa decyzyjnego Netezza	103
Modele użytkowe K-średnie Netezza	104
Modele użytkowe Sieci Bayesa Netezza	104
Modele użytkowe Netezza Naive Bayes	105
Modele użytkowe KNN Netezza	106
Modele użytkowe grupowania dzielącego Netezza	107
Modele użytkowe PCA Netezza	107
Modele użytkowe drzewa regresji Netezza	108
Modele użytkowe regresji liniowej Netezza	109
Model użytkowy szeregów czasowych Netezza	109
Użytkowy uogólniony model liniowy Netezza	110
Model użytkowy Dwustopniowa Netezza	111

Rozdział 7. Modelowanie w bazie danych za pomocą produktu IBM DB2 for z/OS 113

IBM SPSS Modeler and IBM DB2 for z/OS	113
Wymagania dotyczące integracji z produktem IBM DB2 for z/OS	113
Aktywacja integracji z produktem IBM DB2 Analytics Accelerator for z/OS	113
Konfigurowanie programu IBM DB2 for z/OS i IBM Analytics Accelerator for z/OS	114
Tworzenie źródła ODBC dla produktów IBM DB2 for z/OS oraz IBM DB2 Analytics Accelerator	114
Aktywacja integracji produktu IBM DB2 for z/OS w produkcie IBM SPSS Modeler	114
Włączenie opcji generowania i optymalizacji kodu SQL	114
Budowanie modeli za pomocą produktu IBM DB2 for z/OS	115

Modele IBM DB2 for z/OS — Opcje zmiennych	116
Modele IBM DB2 for z/OS — Opcje serwera	116
Modele IBM DB2 for z/OS — Opcje modelu	116
Modele IBM DB2 for z/OS — K-średnie	117
Modele IBM DB2 for z/OS — Opcje zmiennej K-średnie	117
Modele IBM DB2 for z/OS — Opcje budowania zmiennej K-średnie	117
Modele IBM DB2 for z/OS — Naive Bayes	118
Modele IBM DB2 for z/OS — Drzewa decyzyjne	118
Modele IBM DB2 for z/OS — Opcje zmiennych drzewa decyzyjnego	118
Modele IBM DB2 for z/OS — Opcje budowania drzewa decyzyjnego	119
Modele IBM DB2 for z/OS — Węzeł Drzewo decyzyjne — Wagi klas	119
Modele IBM DB2 for z/OS — Węzeł Drzewo decyzyjne — Przycinanie drzewa	120
Modele IBM DB2 for z/OS — Drzewo regresji	120
Modele IBM DB2 for z/OS — Opcje budowania drzewa regresji — Wzrost drzewa	120
Modele IBM DB2 for z/OS — Opcje budowania drzewa regresji — Przycinanie drzewa	121
Modele IBM DB2 for z/OS — Dwustopniowa	121
Modele IBM DB2 for z/OS — Opcje zmiennej Dwustopniowa	122
Modele IBM DB2 for z/OS — Opcje budowania zmiennej Dwustopniowa	122
Modele IBM DB2 for z/OS — Model użytkowy Dwustopniowa — Karta Model	123
Zarządzanie modelami IBM DB2 for z/OS	123
Ocenianie modeli IBM DB2 for z/OS	123
Modele użytkowe drzewa decyzyjnego IBM DB2 for z/OS	123
Model użytkowy K-średnie IBM DB2 for z/OS	124
Modele użytkowe Naive Bayes IBM DB2 for z/OS	124
Model użytkowy drzewa regresji IBM DB2 for z/OS	124
Model użytkowy Dwustopniowa IBM DB2 for z/OS	125

Uwagi. 127
Znaki towarowe 128
Warunki dotyczące dokumentacji produktu 129

Indeks 131

Przedmowa

IBM® SPSS Modeler to oferowane przez IBM Corp. zaawansowane środowisko eksploracji danych. SPSS Modeler pomaga przedsiębiorstwom i instytucjom w rozwijaniu relacji z klientami i obywatelami w oparciu o pogłębioną interpretację dostępnych danych. Organizacje korzystają z wiedzy uzyskanej dzięki programowi SPSS Modeler w bardzo szerokim spektrum zastosowań, m.in. do zatrzymywania najbardziej wartościowych klientów, określania możliwości sprzedaży wiązanej, przyciągania nowych klientów, wykrywania oszustw, ograniczania ryzyka i podnoszenia jakości usług publicznych.

Interfejs graficzny produktu SPSS Modeler zachęca użytkowników, aby wykorzystywali specjalistyczną wiedzę, dzięki której możliwe będzie opracowanie bardziej wydajnych modeli predykcyjnych i skrócenie czasu potrzebnego do uzyskania rozwiązania. SPSS Modeler oferuje wiele technik modelowania, takich jak predykcja, klasyfikacja, segmentacja i algorytmy do wykrywania związków. Po utworzeniu modeli program IBM SPSS Modeler Solution Publisher umożliwia udostępnienie ich osobom podejmującym decyzje w całym przedsiębiorstwie lub zapisanie w bazie danych.

Informacje o programie IBM Business Analytics

Oprogramowanie IBM Business Analytics dostarcza kompletne, spójne i dokładne informacje, na których mogą polegać osoby decyzyjne chcące polepszyć wyniki biznesowe. Wszechstronne portfolio obejmujące moduły: analiza biznesowa, analiza prognostyczna, zarządzanie wynikami i strategiami finansowymi oraz aplikacje analityczne, zapewnia jasny, natychmiastowy i pozwalający na podjęcie działań wgląd w bieżące wyniki oraz daje możliwość przewidywania przyszłych wyników. W połączeniu z licznymi rozwiązaniami branżowymi, sprawdzonymi praktykami i profesjonalnymi usługami, organizacje o różnych rozmiarach mogą wspomagać najwyższą produktywność, w sposób pewny zautomatyzować decyzje i uzyskać lepsze wyniki.

Oprogramowanie IBM SPSS Predictive Analytics będące częścią tego portfolio wspomaga organizacje w zakresie przewidywania przyszłych zdarzeń oraz proaktywnie wpływać na na ten wgląd z korzyścią dla wyników finansowych. Klienci komercyjni, rządowi i uczelnie na całym świecie polegają na technologii IBM SPSS zapewniającej przewagę konkurencyjną, dzięki której przyciągają, zatrzymują i pozyskują nowych klientów, walcząc z nieuczciwością i ograniczając ryzyko. Wdrażając oprogramowanie IBM SPSS do swojej codziennej działalności, organizacje stają się przewidującymi przedsiębiorstwami, zdolnymi do zarządzania i automatyzacji decyzji w celu realizacji celów biznesowych i osiągnięcia mierzalnej przewagi konkurencyjnej. W celu uzyskania dalszych informacji lub skontaktowania się z przedstawicielem, proszę wejść na stronę <http://www.ibm.com/spss>.

Wsparcie techniczne

Wsparcie techniczne jest dostępne w celu zapewnienia klientom obsługi technicznej. Klienci mogą się kontaktować z działem Wsparcia technicznego w celu uzyskania pomocy dotyczącej korzystania z IBM Corp. produktów lub pomocy w instalacji dla jednego z obsługiwanych środowisk sprzętowych. Aby skontaktować się z działem Wsparcia technicznego, wejdź na stronę internetową IBM Corp. pod adresem <http://www.ibm.com/support>. W przypadku prośby o pomoc, należy przygotować swoje dane identyfikacyjne, dane swojej organizacji, a także dane dotyczące usług wsparcia.

Rozdział 1. O programie IBM SPSS Modeler

IBM SPSS Modeler to zestaw narzędzi do eksploracji danych. Produkt umożliwia szybkie opracowywanie modeli predykcyjnych przy wykorzystaniu wiedzy specjalistycznej i stosowanie tych modeli w procesach biznesowych jako wsparcia przy podejmowaniu decyzji. Rozwiązania zawarte w oprogramowaniu IBM SPSS Modeler zapewniają możliwość wykorzystywania branżowego modelu CRISP-DM i pozwalają na obsługę całego procesu eksploracji danych: od pozyskiwania danych do uzyskiwania lepszych wyników biznesowych.

Oprogramowanie IBM SPSS Modeler umożliwia korzystanie z wielu metod modelowania opartych na sztucznej inteligencji, uczeniu maszynowym i statystykach. Metody dostępne na palecie Modelowanie pozwalają na ekstrakowanie nowych informacji z danych i tworzenie modeli predykcyjnych. Każda metoda ma określone mocne strony i jest dostosowana do rozwiązywania określonych problemów.

Oprogramowanie SPSS Modeler można zakupić jako produkt samodzielny lub jako program kliencki używany wraz z oprogramowaniem SPSS Modeler Server. Dostępnych jest wiele opcji dodatkowych, które przedstawiono w kolejnych rozdziałach. Aby uzyskać więcej informacji, patrz <http://www.ibm.com/software/analytics/spss/products/modeler/>.

Produkty IBM SPSS Modeler

Rodzina produktów IBM SPSS Modeler i towarzyszącego im oprogramowania składa się z elementów przedstawionych poniżej.

- IBM SPSS Modeler
- IBM SPSS Modeler Server
- IBM SPSS Modeler Administration Console
- IBM SPSS Modeler Batch
- IBM SPSS Modeler Solution Publisher
- IBM SPSS Modeler Server adapters for IBM SPSS Collaboration and Deployment Services

IBM SPSS Modeler

Oprogramowanie SPSS Modeler to w pełni funkcjonalna wersja produktu instalowana i uruchamiana na komputerze osobistym. Oprogramowanie SPSS Modeler można uruchomić lokalnie jako produkt samodzielny lub korzystać z niego w trybie rozproszonym wraz z serwerem IBM SPSS Modeler Server. Tego typu rozwiązanie zapewnia zwiększenie wydajności obsługi dużych zbiorów danych.

Dzięki oprogramowaniu SPSS Modeler można szybko tworzyć dokładne modele predykcyjne, stosując intuicyjne metody niewymagające umiejętności programowania. Unikatowy interfejs graficzny pozwala na wizualizowanie procedur eksploracji danych. Zaawansowane metody opracowywania analiz dostępne w programie umożliwiają określanie wcześniej niezauważalnych wzorców i trendów zawartych w danych. Użytkownik może modelować wyniki i poznawać czynniki wpływające na ich wartości. W ten sposób można wykorzystywać nowe szanse biznesowe i obniżać ryzyko.

Dostępne są dwie edycje oprogramowania SPSS Modeler: SPSS Modeler Professional oraz SPSS Modeler Premium. Więcej informacji można znaleźć w temacie “Wydania programu IBM SPSS Modeler” na stronie 2.

IBM SPSS Modeler Server

Oprogramowanie SPSS Modeler działa w oparciu o architekturę klient-serwer, w której żądania wymagające zaangażowania dużych zasobów kierowane są do zaawansowanego oprogramowania serwerowego. Takie rozwiązanie umożliwia bardziej wydajną obsługę dużych zbiorów danych.

SPSS Modeler Server to produkt wymagający dodatkowej licencji, działający stale na serwerze w trybie analizy rozproszonej. Współpracuje on z co najmniej jedną instalacją oprogramowania IBM SPSS Modeler. W ten sposób oprogramowanie SPSS Modeler Server poprawia wydajność podczas obsługi dużych zbiorów danych, ponieważ operacje wymagające dużej mocy obliczeniowej można wykonywać na serwerze bez potrzeby pobierania danych na komputer kliencki. Oprogramowanie IBM SPSS Modeler Server optymalizuje również obsługę SQL i funkcje modelowania wewnątrz bazy danych, co dodatkowo zwiększa wydajność działania i sprzyja automatyzacji pracy.

IBM SPSS Modeler Administration Console

Oprogramowanie Modeler Administration Console to aplikacja graficzna służąca do obsługi wielu opcji konfiguracji SPSS Modeler Server, które można dostosować również za pomocą pliku opcji. W ramach aplikacji dostępny jest interfejs konsoli użytkownika pozwalający na monitorowanie i konfigurowanie instalacji SPSS Modeler Server. Interfejs jest dostępny bez dodatkowych opłat dla aktualnych użytkowników SPSS Modeler Server. Aplikację można zainstalować tylko na komputerach z systemem Windows, jednak administrować można serwerem zainstalowanym na dowolnej obsługiwanej platformie.

IBM SPSS Modeler Batch

Eksploatacja danych jest zazwyczaj procesem interaktywnym, jednak oprogramowanie SPSS Modeler można też uruchomić z poziomu wiersza komend i zrezygnować z używania graficznego interfejsu użytkownika. Niekiedy użytkownik wykonuje długotrwałe lub powtarzalne zadania, które mogą być realizowane bez nadzoru.

Oprogramowanie SPSS Modeler Batch to specjalna wersja produktu pozwalająca na wykonywanie wszystkich funkcji analitycznych SPSS Modeler bez potrzeby używania standardowego interfejsu użytkownika. Do korzystania z aplikacji SPSS Modeler Batch wymagane jest oprogramowanie SPSS Modeler Server.

IBM SPSS Modeler Solution Publisher

SPSS Modeler Solution Publisher umożliwia tworzenie spakowanych wersji strumieni programu SPSS Modeler, które można uruchamiać za pomocą zewnętrznych środowisk wykonawczych lub osadzać w aplikacji zewnętrznej. W ten sposób można publikować i wdrażać pełne strumienie SPSS Modeler w celu używania ich w środowiskach, w których nie zainstalowano programu SPSS Modeler. SPSS Modeler Solution Publisher jest dystrybuowany jako część produktu IBM SPSS Collaboration and Deployment Services - Scoring, który do działania wymaga oddzielnej licencji. Wraz z licencją użytkownik otrzymuje oprogramowanie SPSS Modeler Solution Publisher Runtime umożliwiające uruchamianie opublikowanych strumieni.

Więcej informacji na temat programu SPSS Modeler Solution Publisher znajduje się w dokumentacji produktu IBM SPSS Collaboration and Deployment Services. W Centrum wiedzy IBM SPSS Collaboration and Deployment Services dostępne są sekcje "IBM SPSS Modeler Solution Publisher" oraz "IBM SPSS Analytics Toolkit".

IBM SPSS Modeler Server Adapters for IBM SPSS Collaboration and Deployment Services

Dostępnych jest wiele adapterów dla IBM SPSS Collaboration and Deployment Services, które umożliwiają współpracę programów SPSS Modeler i SPSS Modeler Server z repozytorium IBM SPSS Collaboration and Deployment Services. Dzięki temu strumień SPSS Modeler wdrożony w repozytorium można udostępnić wielu użytkownikom lub uzyskać do niego dostęp z poziomu uproszczonej aplikacji klienckiej IBM SPSS Modeler Advantage. Adapter należy zainstalować na systemie hostującym repozytorium.

Wydania programu IBM SPSS Modeler

Dostępne są następujące wydania oprogramowania SPSS Modeler.

SPSS Modeler Professional

Oprogramowanie SPSS Modeler Professional zapewnia wszystkie narzędzia wymagane do obsługi większości typów danych ustrukturyzowanych, takich jak np. zachowania i interakcje śledzone w systemach CRM, dane demograficzne, zachowania zakupowe i dane sprzedażowe.

SPSS Modeler Premium

Oprogramowanie SPSS Modeler Premium wymaga oddzielnej licencji. Dzięki temu rozwiązaniu oprogramowanie SPSS Modeler Professional może obsługiwać wyspecjalizowane dane, np. używane z technikami Entity Analytics lub analizami sieci społecznościowych, oraz nieustrukturyzowane dane tekstowe. Oprogramowanie SPSS Modeler Premium składa się z następujących komponentów.

IBM SPSS Modeler Entity Analytics dodaje dodatkowy wymiar do analiz predykcyjnych wykonywanych przy użyciu produktu IBM SPSS Modeler. Analizy predykcyjne stanowią próbę przewidzenia przyszłych zachowań na podstawie danych z przeszłości, natomiast techniki Entity Analytics służą przede wszystkim poprawie spójności danych bieżących poprzez rozwiązywanie konfliktów dotyczących tożsamości w samych rekordach. Tożsamość może dotyczyć osoby, organizacji, obiektu lub dowolnej innej jednostki, względem której może istnieć niejednoznaczność. Ustalenie tożsamości może być istotne w przypadku wielu dziedzin, takich jak zarządzanie relacjami z klientami, wykrywanie oszustw, pranie pieniędzy oraz bezpieczeństwo w skali krajowej i międzynarodowej.

Oprogramowanie **IBM SPSS Modeler Social Network Analysis** przetwarza informacje o relacjach w zmienne, które charakteryzują zachowania społeczne pojedynczych osób i grup. Korzystając z danych opisujących relacje w sieciach społecznościowych, oprogramowanie IBM SPSS Modeler Social Network Analysis identyfikuje liderów społecznych, którzy wpływają na zachowanie innych w sieci. Ponadto umożliwia ustalenie ludzi, którzy są pod największym wpływem innych uczestników sieci. Łącząc te wyniki z innymi środkami, można stworzyć obszerne profile osób, na których będą bazowały modele predykcyjne użytkownika. Modele zawierające te informacje społeczne będą dawały lepsze wyniki niż modele, które ich nie zawierają.

Program **IBM SPSS Modeler Text Analytics** korzysta z zaawansowanych rozwiązań lingwistycznych oraz przetwarzania języka naturalnego w celu szybkiego przetwarzania różnego rodzaju nieustrukturyzowanych danych tekstowych, wyodrębniania i porządkowania kluczowych pojęć oraz grupowania tych pojęć w kategorie. Wyodrębnione pojęcia i kategorie można łączyć z istniejącymi danymi ustrukturyzowanymi, takimi jak dane demograficzne, a następnie stosować w celu modelowania, korzystając z produktu IBM SPSS Modeler i zawartego w nim pełnego pakietu narzędzi do eksploracji danych, aby w rezultacie takiego połączenia podejmować lepsze decyzje przy zmniejszonej ilości zakłóceń.

Dokumentacja IBM SPSS Modeler

Dokumentacja elektroniczna jest dostępna w programie SPSS Modeler z poziomu menu Pomoc. Zasoby te obejmują informacje na temat oprogramowania SPSS Modeler, SPSS Modeler Server i publikację Applications — podręcznik (nazywaną również Samouczkiem) oraz inne materiały pomocnicze.

Pełna dokumentacja dla każdego produktu (obejmująca instrukcje instalacji) jest dostępna w formacie PDF, w osobnym, skompresowanym folderze, w ramach materiałów dotyczących produktu do pobrania. Dokumenty dotyczące instalacji można również pobrać z Internetu pod adresem <http://www.ibm.com/support/docview.wss?uid=swg27046871>.

Dokumentacja w obu formatach jest również dostępna w Centrum Wiedzy SPSS Modeler pod adresem http://www-01.ibm.com/support/knowledgecenter/SS3RA7_18.1.0.

Dokumentacja SPSS Modeler Professional

Pakiet dokumentacji SPSS Modeler Professional (bez instrukcji instalacyjnych) zawiera następujące publikacje.

- **IBM SPSS Modeler — podręcznik użytkownika.** Ogólne wprowadzenie do obsługi oprogramowania SPSS Modeler, obejmuje opisy procedur tworzenia strumieni danych, obsługi braków danych, tworzenia wyrażeń CLEM, pracy z projektami i raportami, dane na temat przygotowywania strumieni z przeznaczeniem do wdrażania w IBM SPSS Collaboration and Deployment Services lub IBM SPSS Modeler Advantage.
- **IBM SPSS Modeler — węzły źródłowe, procesowe i wyników.** Opisy wszystkich węzłów używanych do odczytywania, przetwarzania i tworzenia wynikowych postaci danych w różnych formatach. Czyli wszystkich węzłów poza węzłami modelowania.

- **IBM SPSS Modeler — węzły modelowania.** Opisy wszystkich węzłów używanych do tworzenia modeli eksploracji danych. Oprogramowanie IBM SPSS Modeler umożliwia korzystanie z wielu metod modelowania opartych na sztucznej inteligencji, uczeniu maszynowym i statystykach.
- **IBM SPSS Modeler Algorithms Guide.** Opisy podstaw matematycznych dotyczących metod modelowania stosowanych w oprogramowaniu IBM SPSS Modeler. Ten podręcznik jest dostępny tylko w formacie PDF.
- **IBM SPSS Modeler — podręcznik zastosowań.** Przykłady zawarte w niniejszym przewodniku stanowią skrócone informacje związane z konkretnymi metodami i technikami modelowania. Wersja online tego podręcznika jest również dostępna z poziomu menu Pomoc. Więcej informacji można znaleźć w temacie “Przykłady aplikacji”.
- **IBM SPSS Modeler — podręcznik tworzenia skryptów w języku Python i automatyzacji.** Informacje na temat automatyzacji działania systemu za pomocą skryptów Python wraz z właściwościami służącymi do obsługi węzłów i strumieni.
- **IBM SPSS Modeler — podręcznik wdrażania.** Informacje na temat uruchamiania strumieni IBM SPSS Modeler przedstawione w postaci krokowych operacji wykonywanych podczas przetwarzania zadań w oprogramowaniu IBM SPSS Collaboration and Deployment Services Deployment Manager.
- **IBM SPSS Modeler CLEF Developer's Guide.** Z oprogramowaniem CLEF można zintegrować inne programy pozwalające na przetwarzanie danych lub obsługę algorytmów modelujących w postaci węzłów w IBM SPSS Modeler.
- **IBM SPSS Modeler — podręcznik eksploracji w bazie danych.** Informacje na temat wydajnego wykorzystywania bazy danych w celu zwiększenia wydajności i zakresu funkcji analitycznych za pomocą algorytmów innych firm.
- **IBM SPSS Modeler Server — podręcznik administracji i wydajności.** Informacje na temat konfiguracji i funkcji administracyjnych w oprogramowaniu IBM SPSS Modeler Server.
- **IBM SPSS Modeler Administration Console — podręcznik użytkownika.** Informacje na temat instalowania i używania interfejsu użytkownika konsoli w celu monitorowania i konfigurowania oprogramowania IBM SPSS Modeler Server. Konsola działa jako wtyczka do aplikacji Deployment Manager.
- **IBM SPSS Modeler CRISP-DM — podręcznik.** Szczegółowy podręcznik metodologii CRISP-DM w kontekście eksploracji danych za pomocą oprogramowania SPSS Modeler.
- **IBM SPSS Modeler Batch — podręcznik użytkownika.** Pełny podręcznik obsługi oprogramowania IBM SPSS Modeler w trybie wsadowym obejmujący szczegółowe informacje na temat pracy w trybie wsadowym i korzystania z argumentów z poziomu wiersza komend. Ten podręcznik jest dostępny tylko w formacie PDF.

Dokumentacja SPSS Modeler Premium

Pakiet dokumentacji produktu SPSS Modeler Premium (bez instrukcji instalacyjnych) zawiera następujące publikacje.

- **IBM SPSS Modeler Entity Analytics — podręcznik użytkownika.** Informacje dotyczące używania techniki Entity Analytics w oprogramowaniu SPSS Modeler obejmują procedury dotyczące instalacji i konfiguracji repozytorium, węzłów technik Entity Analytics oraz zadań administracyjnych.
- **IBM SPSS Modeler Social Network Analysis — podręcznik użytkownika .** Podręcznik zawierający informacje na temat analizy sieci społecznościowych za pomocą oprogramowania SPSS Modeler w tym informacje o analizie grup i analizie przenikania.
- **SPSS Modeler Text Analytics — podręcznik użytkownika.** Informacje na temat używania analiz tekstu za pomocą oprogramowania SPSS Modeler, obejmują procedury dotyczące węzłów eksploracji tekstu, interaktywnego pulpitu roboczego, szablonów oraz innych zasobów.

Przykłady aplikacji

Podczas gdy narzędzia do eksploracji danych w programie SPSS Modeler mogą pomóc w rozwiązaniu szeregu problemów biznesowych i organizacyjnych, przykłady aplikacji udostępniają krótkie, ukierunkowane wprowadzenia do konkretnych metod i technik modelowania. Używane tutaj zestawy danych są znacznie mniejsze niż ogromne składnice danych zarządzane przez programy do eksploracji danych, lecz używane koncepcje i metody są skalowalne odpowiednio do potrzeb rzeczywistych aplikacji.

Dostęp do przykładów można uzyskać, klikając opcję **Przykłady aplikacji** w menu Pomoc programu SPSS Modeler.

Pliki danych i przykładowe strumienie są instalowane w folderze **Dema**, w katalogu instalacyjnym produktu. Aby uzyskać więcej informacji, patrz “Folder Demos”.

Przykłady modelowania w bazach danych. Przykłady zamieszczono w publikacji *IBM SPSS Modeler — podręcznik eksploracji w bazie danych*.

Przykłady skryptów. Przykłady zamieszczono w publikacji *IBM SPSS Modeler — podręcznik tworzenia skryptów w języku Python i automatyzacji*.

Folder Demos

Pliki danych i przykładowe strumienie używane z przykładami do aplikacji są instalowane w folderze **Demos** wewnątrz katalogu instalacyjnego produktu (na przykład: C:\Program Files\IBM\SPSS\Modeler\\Demos). Dostęp do tego folderu można także uzyskać z grupy programów IBM SPSS Modeler w menu Start systemu Windows lub klikając opcję **Demos** na liście ostatnich katalogów w oknie dialogowym **Plik > Otwórz strumień**.

Monitorowanie wykorzystania licencji

Podczas pracy z produktem SPSS Modeler wykorzystanie licencji jest monitorowane i regularnie rejestrowane. Metryka wykorzystania licencji nosi nazwę *AUTHORIZED_USER* (użytkownik autoryzowany) lub *CONCURRENT_USER* (użytkownik pracujący jednocześnie), a typ rejestrowanej metryki zależy od typu licencji na produkt SPSS Modeler, którą posiada użytkownik.

Generowane pliki dzienników mogą być przetwarzane przez program IBM License Metric Tool, z którego uzyskać można raporty o wykorzystaniu licencji.

Pliki dzienników wykorzystania licencji są tworzone w tym samym katalogu, w którym zapisywane są dzienniki klienta SPSS Modeler (domyślnie %ALLUSERSPROFILE%/IBM/SPSS/Modeler/<wersja>/log).

Rozdział 2. Eksploracja w bazie danych

Przegląd informacji o modelowaniu w bazie danych

Program IBM SPSS Modeler Server obsługuje integrację z narzędziami do eksploracji danych i modelowania, które są dostępne w bazach danych, takich jak IBM Netezza, IBM DB2 InfoSphere Warehouse, Oracle Data Miner i Microsoft Analysis Services. Można tworzyć, oceniać i składować modele w bazie danych — a wszystko to w ramach aplikacji IBM SPSS Modeler. Dzięki temu można połączyć możliwości analityczne i łatwość użycia produktu IBM SPSS Modeler z mocą i wydajnością bazy danych, jednocześnie czerpiąc korzyści z algorytmów rodzimych dla bazy danych, które są udostępniane przez tych dostawców. Modele są budowane wewnątrz bazy danych, którą następnie można przeglądać i oceniać przez interfejs IBM SPSS Modeler w normalny sposób, a także w razie potrzeby wdrażać, korzystając z produktu IBM SPSS Modeler Solution Publisher. Obsługiwane algorytmy znajdują się na palecie Modelowanie w bazie w produkcie IBM SPSS Modeler.

Korzystanie z produktu IBM SPSS Modeler w celu uzyskiwania dostępu do algorytmów rodzimych dla bazy danych ma kilka zalet:

- Algorytmy w bazie danych są często ściśle zintegrowane z serwerem bazy danych i mogą zapewniać poprawę wydajności.
- Modele budowane i zapisywane „w bazie danych” mogą być łatwiej wdrażane do wszelkich aplikacji, które mają dostęp do bazy danych, a także udostępniane tym aplikacjom.

Generowanie kodu SQL. Modelowanie w bazie danych przebiega odrębnie od generowania kodu SQL określonego również jako „analiza wstępna SQL”. Ta funkcja umożliwi generowanie instrukcji SQL dla rodzimych operacji produktu IBM SPSS Modeler, które mogą być „wstawiane” (czyli wykonywane) w bazie danych w celu poprawy wydajności. Węzeł łączenia, węzeł agregacji i węzeł selekcji generują kod SQL, który może zostać w ten sposób wstawiony do bazy danych. Generowanie kodu SQL w połączeniu z modelowaniem w bazie danych może wywołać strumienie, które mogą być od początku do końca wykonywane w bazie danych, co powoduje znaczącą poprawę wydajności w porównaniu ze strumieniami wykonywanymi w produkcie IBM SPSS Modeler.

Uwaga: Modelowanie w bazie danych i optymalizacja SQL wymagają włączenia na komputerze z programem IBM SPSS Modeler możliwości połączenia z serwerem IBM SPSS Modeler Server. Po włączeniu tej opcji można uzyskać dostęp do algorytmów baz danych, wstawić SQL do kolejki bezpośrednio z programu IBM SPSS Modeler i uzyskać dostęp do programu IBM SPSS Modeler Server. W celu sprawdzenia bieżącego statusu licencji należy wybrać z menu programu IBM SPSS Modeler następujące opcje.

Pomoc > Informacje o programie > Dodatkowe szczegóły

Po włączeniu możliwości połączenia na karcie Status licencji widoczna jest opcja **Aktywacja serwera**.

Informacje dotyczące obsługiwanych algorytmów zawierają kolejne sekcje dotyczące poszczególnych dostawców.

Co jest potrzebne

W celu przeprowadzenia modelowania w bazie danych wymagana jest następująca konfiguracja:

- Połączenie ODBC do odpowiedniej bazy danych z zainstalowanymi wymaganymi składnikami analitycznymi (Microsoft Analysis Services, Oracle Data Miner lub IBM DB2 InfoSphere Warehouse).
- W produkcie IBM SPSS Modeler modelowanie w bazie danych należy włączyć w oknie dialogowym Aplikacje pomocnicze (**Narzędzia > Aplikacje pomocnicze**).
- Ustawienia **Generuj kod SQL kierowany do bazy** i **Optymalizacja SQL** należy włączyć w oknie dialogowym Opcje użytkownika w produkcie IBM SPSS Modeler, a także w produkcie IBM SPSS Modeler Server (jeśli jest używany). Należy zwrócić uwagę na to, że optymalizacja SQL nie jest ściśle wymagana do tego, aby modelowanie w bazie danych działało, ale jest bardzo zalecane ze względu na wydajność.

Uwaga: Modelowanie w bazie danych i optymalizacja SQL wymagają włączenia na komputerze z programem IBM SPSS Modeler możliwości połączenia z serwerem IBM SPSS Modeler Server. Po włączeniu tej opcji można uzyskać dostęp do algorytmów baz danych, wstawić SQL do kolejki bezpośrednio z programu IBM SPSS Modeler i uzyskać dostęp do programu IBM SPSS Modeler Server. W celu sprawdzenia bieżącego statusu licencji należy wybrać z menu programu IBM SPSS Modeler następujące opcje.

Pomoc > Informacje o programie > Dodatkowe szczegóły

Po włączeniu możliwości połączenia na karcie Status licencji widoczna jest opcja **Aktywacja serwera**.

Szczegółowe informacje zawierają kolejne sekcje dotyczące poszczególnych dostawców.

Budowanie modelu

Proces budowania i oceniania modeli przy użyciu algorytmów bazodanowych jest podobny do innych rodzajów eksploracji danych w produkcie IBM SPSS Modeler. Ogólny proces pracy z węzłami i „modelami użytkowymi” modelowania jest podobny do dowolnego innego strumienia stosowanego w przypadku pracy z produktem IBM SPSS Modeler. Jedyną różnicą polega na tym, że rzeczywiste operacje przetwarzania i budowania modelu są wstawiane do bazy danych.

Strumień modelowania bazy danych jest koncepcyjnie identyczny z innymi strumieniami danych w produkcie IBM SPSS Modeler; jednak ten strumień wykonuje wszystkie operacje w bazie danych, w tym także budowanie modelu z użyciem węzła MS Decision Tree. Po uruchomieniu strumienia produkt IBM SPSS Modeler przekazuje do bazy danych instrukcję budowania i przechowywania wynikowego modelu, a wyniki są pobierane do produktu IBM SPSS Modeler. Wykonanie w bazie danych jest wskazywane przez użycie w strumieniu fioletowych węzłów.

Przygotowanie danych

Bez względu na to, czy algorytmy rodzime bazy danych są używane, czy nie, przygotowania danych powinny być wstawiane do bazy danych, gdy tylko jest to możliwe, aby poprawić wydajność.

- Jeśli dane oryginalne są zapisywane w bazie danych, wówczas celem jest zachowanie ich tam poprzez upewnienie się, że wszystkie wymagane poprzedzające operacje mogą zostać przekształcone na kod SQL. Takie postępowanie zapobiega pobieraniu danych do produktu IBM SPSS Modeler — w ten sposób można uniknąć wąskiego gardła, które może anulować wszelkie korzyści — i umożliwia uruchomienie całego strumienia w bazie danych.
- Jeśli dane oryginalne *nie* są zapisywane w bazie danych, nadal możliwe jest korzystanie z modelowania w bazie danych. W takim przypadku przygotowanie danych jest realizowane w produkcie IBM SPSS Modeler, a przygotowany zestaw danych jest automatycznie wczytywany do bazy danych na potrzeby budowania modelu.

Ocenianie modelu

Modele wygenerowane z produktu IBM SPSS Modeler przy użyciu eksploracji w bazie danych różnią się od standardowych modeli IBM SPSS Modeler. Mimo że są widoczne w menedżerze modeli jako modele użytkowe wygenerowanego modelu, w rzeczywistości są zdalnymi modelami, które znajdują się na zdalnym serwerze eksploracji danych lub zdalnym serwerze bazy danych. W produkcie IBM SPSS Modeler widoczne są tylko proste odniesienia do tych zdalnych modeli. Innymi słowy, widoczny model IBM SPSS Modeler jest pustym modelem, który zawiera informacje, takie jak nazwa hosta serwera bazy danych, nazwa bazy danych i nazwa modelu. Jest to istotna różnica, którą należy zrozumieć podczas przeglądania i oceny modeli utworzonych przy użyciu algorytmów rodzimych dla bazy danych.

Po utworzeniu modelu można go dodać do strumienia w celu oceny, podobnie jak każdym inny model wygenerowany w produkcie IBM SPSS Modeler. Wszystkie operacje oceny są realizowane w bazie danych, nawet jeśli poprzedzające operacje nie były tam wykonywane. (Operacje poprzedzające nadal mogą być wstawiane do bazy danych, jeśli możliwa jest poprawa wydajności, jednak nie jest to wymagane w celu realizacji oceniania). Ponadto w większości przypadków wygenerowany model można przeglądać przy użyciu standardowej przeglądarki udostępnianej przez dostawcę bazy danych.

W celu przeglądania i oceniania wymagane jest aktywne połączenie z serwerem, na którym uruchomiony jest produkt Oracle Data Miner, IBM DB2 InfoSphere Warehouse lub Microsoft Analysis Services.

Wyświetlanie wyników i określanie ustawień

W celu wyświetlania wyników i określania ustawień oceniania kliknij dwukrotnie model w obszarze roboczym strumienia. Alternatywnie można również kliknąć prawym przyciskiem myszy model i wybrać opcję **Przełączaj** lub **Edytuj**. Konkretnie ustawienia są zależne od typu modelu.

Eksportowanie i zapisywanie modeli bazodanowych

Modele bazodanowe i podsumowania modeli mogą być eksportowane z przeglądarki modeli w taki sam sposób, jak inne modele utworzone w produkcie IBM SPSS Modeler, czyli przy użyciu opcji z menu Plik.

1. Z menu Plik w przeglądarce modeli wybierz dowolną z poniższych opcji:

- **Eksportuj do pliku tekstowego** — umożliwia wyeksportowanie podsumowania modelu do pliku tekstowego
- **Eksportuj do HTML** — umożliwia wyeksportowanie podsumowania modelu do pliku HTML
- **Eksportuj PMML** (obsługiwane tylko w przypadku modeli IBM DB2 IM) — umożliwia wyeksportowanie modelu w formacie języka PMML (Predictive Model Markup Language), który może zostać użyty z innym oprogramowaniem obsługującym format języka PMML.

Uwaga: wygenerowany model można również zapisać poprzez wybranie opcji **Zapisz węzeł** z menu Plik.

Spójność modelu

W przypadku każdego wygenerowanego modelu bazy danych produkt IBM SPSS Modeler zapisuje opis struktury modelu razem z odniesieniem do modelu o tej samej nazwie, która jest zapisana w bazie danych. Na karcie Serwer wygenerowanego modelu wyświetlany jest unikalny klucz wygenerowany dla tego modelu, który jest zgodny z rzeczywistym modelem w bazie danych.

Produkt IBM SPSS Modeler wykorzystuje ten losowo wygenerowany klucz, aby sprawdzić, czy model jest nadal spójny. Klucz jest zapisywany w opisie modelu podczas jego budowania. Przed uruchomieniem strumienia wdrażania warto sprawdzić, czy klucze są zgodne.

1. W celu sprawdzenia spójności modelu zapisanego w bazie danych poprzez porównanie jego opisu z losowym kluczem zapisanym przez produkt IBM SPSS Modeler należy kliknąć przycisk **Sprawdź**. Jeśli modelu bazy danych nie można znaleźć lub klucz jest niezgodny, zostanie zgłoszony błąd.

Wyświetlanie i eksportowanie wygenerowanego kodu SQL

Wygenerowany kod SQL można wyświetlić w postaci podglądu przed jego wykonaniem, co może być pomocne w przypadku debugowania.

Rozdział 3. Modelowanie w bazie danych przy użyciu programu Microsoft Analysis Services

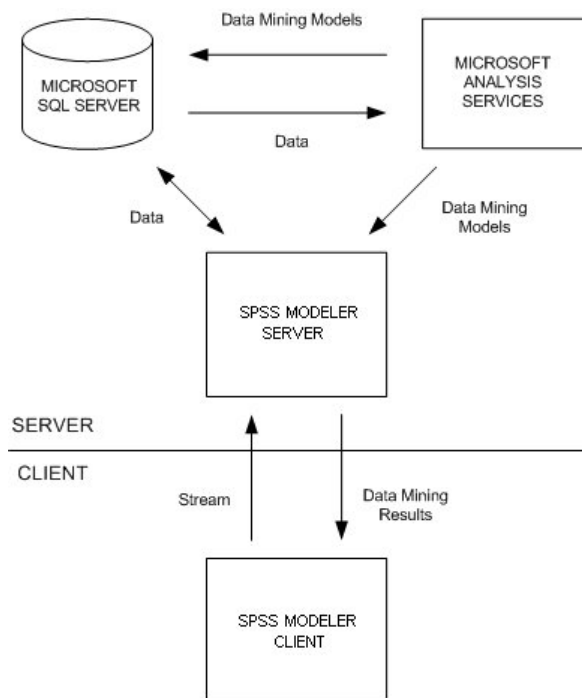
Produkt IBM SPSS Modeler i usługi Microsoft Analysis Services

Produkt IBM SPSS Modeler obsługuje integrację z usługami Microsoft SQL Server Analysis Services. Ta funkcjonalność jest implementowana jako węzły modelowania w produkcie IBM SPSS Modeler i jest dostępna z palety Modelowanie w bazie. Jeśli paleta jest niewidoczna, można ją aktywować, włączając integrację z MS Analysis Services — odpowiednie opcje są dostępne z karty Microsoft w oknie dialogowym Aplikacje pomocnicze. Więcej informacji można znaleźć w temacie “Aktywacja integracji z usługami Analysis Services” na stronie 13.

Produkt IBM SPSS Modeler obsługuje integrację następujących algorytmów usługi Analysis Services:

- Drzewa decyzyjne
- Grupowanie
- Reguły asocjacyjne
- Naive Bayes
- Regresja liniowa
- Sieć neuronowa
- Regresja logistyczna
- Szereg czasowy
- Grupowanie sekwencyjne

Poniższy diagram ilustruje przepływ danych z klienta do serwera, gdzie eksploracja w bazie danych jest zarządzana przez produkt IBM SPSS Modeler Server. Budowanie modelu realizują usługi Analysis Services. Model wynikowy jest zapisywany przez usługi Analysis Services. Odniesienie do tego modelu jest utrzymywane w strumieniach IBM SPSS Modeler. Następnie model jest pobierany z usług Analysis Services do programu Microsoft SQL Server lub produktu IBM SPSS Modeler w celu oceny.



Rysunek 1. Przepływ danych między produktem IBM SPSS Modeler, programem Microsoft SQL Server a usługami Microsoft Analysis Services podczas budowania modelu

Uwaga: produkt IBM SPSS Modeler Server nie jest wymagany, ale może być używany. Klient IBM SPSS Modeler jest zdolny do samodzielnego przetwarzania obliczeń związanych z eksploracją w bazie danych.

Wymagania dotyczące integracji z programem Microsoft Analysis Services

Następujące warunki stanowią warunki wstępne do przeprowadzenia modelowania w bazie danych z użyciem algorytmów Analysis Services w produkcie IBM SPSS Modeler. W celu upewnienia się, że te warunki są spełnione, konieczne mogą być konsultacje z administratorem bazy danych.

- Produkt IBM SPSS Modeler uruchomiony względem instalacji IBM SPSS Modeler Server (tryb rozproszony) w systemie Windows. Platformy UNIX nie są obsługiwane w zakresie tej integracji z użyciem usług Analysis Services.

Ważne: użytkownicy produktu IBM SPSS Modeler muszą skonfigurować połączenie ODBC, używając sterownika SQL Native Client udostępnianego przez firmę Microsoft pod adresem URL podanym poniżej nagłówka *Dodatkowe wymagania dotyczące produktu IBM SPSS Modeler Server*. Sterownik udostępniany z pakietem *IBM SPSS Data Access Pack* (i zwykle polecany do innych zastosowań z produktem IBM SPSS Modeler) nie jest polecany w tym przypadku. Sterownik powinien być skonfigurowany w taki sposób, aby używał programu SQL Server z włączonym **Zintegrowanym uwierzytelnianiem systemu Windows**, ponieważ produkt IBM SPSS Modeler nie obsługuje uwierzytelniania SQL Server. W przypadku pytań dotyczących tworzenia lub określania uprawnień dla źródeł danych ODBC należy skontaktować się z administratorem bazy danych.

- SQL Server musi być zainstalowany, chociaż niekoniecznie na tym samym hoście co IBM SPSS Modeler. Użytkownicy produktu IBM SPSS Modeler muszą mieć uprawnienia wystarczające do odczytu i zapisu danych oraz usuwania i tworzenia tabel oraz widoków.

Uwaga: zalecana jest wersja SQL Server Enterprise Edition. Wersja Enterprise Edition zapewnia dodatkową elastyczność, ponieważ udostępnia rozszerzone parametry, które pozwalają na precyzyjne dostosowywanie wyników działania algorytmów. Wersja Standard Edition zapewnia te same parametry, ale nie pozwala użytkownikom na edytowanie niektórych parametrów zaawansowanych.

- Rozwiązanie Microsoft SQL Server Analysis Services musi być zainstalowane na tym samym hoście, na którym zainstalowany jest program SQL Server.

Dodatkowe wymagania dotyczące produktu IBM SPSS Modeler Server

Korzystanie z algorytmów Analysis Services z produktem IBM SPSS Modeler Server jest możliwe, pod warunkiem że na hoście IBM SPSS Modeler Server zainstalowane będą następujące składniki.

Uwaga: jeśli program SQL Server jest zainstalowany na tym samym hoście, na którym zainstalowany jest produkt IBM SPSS Modeler Server, wówczas te składniki będą już dostępne.

- Microsoft SQL Server Analysis Services 10.0 OLE DB Provider (należy wybrać odpowiedni wariant dla konkretnego systemu operacyjnego)
- Microsoft SQL Server Native Client (należy wybrać odpowiedni wariant dla konkretnego systemu operacyjnego)
- Jeśli używany jest program Microsoft SQL Server 2008, wówczas konieczny może być także produkt Microsoft Core XML Services (MSXML) 6.0.

W celu pobrania tych składników należy odwiedzić stronę www.microsoft.com/downloads, wyszukać **.NET Framework** lub (w przypadku wszystkich pozostałych składników) **SQL Server Feature Pack**, a następnie wybrać najnowszy pakiet dla posiadanej wersji SQL Server.

W przypadku tych składników konieczne może być wcześniejsze zainstalowanie innych pakietów, które również powinny być dostępne na stronie WWW Microsoft Downloads.

Dodatkowe wymagania dotyczące produktu IBM SPSS Modeler

Korzystanie z algorytmów Analysis Services z produktem IBM SPSS Modeler jest możliwe, pod warunkiem że na kliencie zainstalowane będą składniki wymienione powyżej, a także poniższe:

- Microsoft SQL Server Datamining Viewer Controls (należy wybrać odpowiedni wariant dla konkretnego systemu operacyjnego) — w przypadku tego składnika wymagany jest także produkt:
- Microsoft ADOMD.NET

W celu pobrania tych składników należy odwiedzić stronę www.microsoft.com/downloads, wyszukać **SQL Server Feature Pack**, a następnie wybrać najnowszy pakiet dla posiadanej wersji SQL Server.

Uwaga: Modelowanie w bazie danych i optymalizacja SQL wymagają włączenia na komputerze z programem IBM SPSS Modeler możliwości połączenia z serwerem IBM SPSS Modeler Server. Po włączeniu tej opcji można uzyskać dostęp do algorytmów baz danych, wstawić SQL do kolejki bezpośrednio z programu IBM SPSS Modeler i uzyskać dostęp do programu IBM SPSS Modeler Server. W celu sprawdzenia bieżącego statusu licencji należy wybrać z menu programu IBM SPSS Modeler następujące opcje.

Pomoc > Informacje o programie > Dodatkowe szczegóły

Po włączeniu możliwości połączenia na karcie Status licencji widoczna jest opcja **Aktywacja serwera**.

Aktywacja integracji z usługami Analysis Services

Aby włączyć integrację produktu IBM SPSS Modeler z usługami Analysis Services, należy skonfigurować program SQL Server i usługi Analysis Services, utworzyć źródło ODBC, włączyć integrację w oknie dialogowym Aplikacje pomocnicze w produkcie IBM SPSS Modeler, a także włączyć generowanie i optymalizację kodu SQL.

Uwaga: dostępny musi być program Microsoft SQL Server oraz usługi Microsoft Analysis Services. Więcej informacji można znaleźć w temacie “Wymagania dotyczące integracji z programem Microsoft Analysis Services” na stronie 12.

Konfigurowanie programu SQL Server

Skonfiguruj program SQL Server, aby umożliwić ocenianie w bazie danych.

1. Utwórz następujący klucz rejestru na komputerze hoście programu SQL Server:

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

2. Do tego klucza dodaj następującą wartość DWORD:
AllowInProcess 1
3. Zrestartuj program SQL Server po wprowadzeniu tej zmiany.

Konfigurowanie usług Analysis Services

Zanim produkt IBM SPSS Modeler będzie mógł się komunikować z usługami Analysis Services, należy ręcznie skonfigurować dwa ustawienia w oknie dialogowym właściwości usług Analysis Services:

1. Zaloguj się do serwera Analysis Server za pośrednictwem produktu MS SQL Server Management Studio.
2. Uzyskaj dostęp do okna dialogowego właściwości, klikając prawym przyciskiem myszy nazwę serwera i wybierając opcję **Właściwości**.
3. Zaznacz pole wyboru **Pokaż zaawansowane (wszystkie) właściwości**.
4. Zmień następujące właściwości:
 - Zmień wartość DataMining\AllowAdHocOpenRowsetQueries na True (wartością domyślną jest False).
 - Zmień wartość DataMining\AllowProvidersInOpenRowset na [all] (wartość domyślna nie istnieje).

Tworzenie nazwy DSN ODBC dla programu SQL Server

Aby odczytać lub zapisać dane w bazie danych, należy mieć zainstalowane źródło danych ODBC, które jest skonfigurowane dla odpowiedniej bazy danych z uprawnieniami odczytu i zapisu zgodnie z potrzebami. Sterownik Microsoft SQL Native Client ODBC jest wymagany automatycznie instalowany z programem SQL Server. *Sterownik udostępniany z pakietem IBM SPSS Data Access Pack (i zwykle polecany do innych zastosowań z produktem IBM SPSS Modeler) nie jest polecany w tym przypadku.* Jeśli produkt IBM SPSS Modeler i program SQL Server znajdują się na innych hostach, można pobrać sterownik Microsoft SQL Native Client ODBC. Więcej informacji można znaleźć w temacie “Wymagania dotyczące integracji z programem Microsoft Analysis Services” na stronie 12.

W przypadku pytań dotyczących tworzenia lub określania uprawnień dla źródeł danych ODBC należy skontaktować się z administratorem bazy danych.

1. Korzystając ze sterownika Microsoft SQL Native Client ODBC, utwórz nazwę DSN ODBC, która będzie wskazywać na bazę danych programu SQL Server używaną w procesie eksploracji danych. Należy użyć pozostałych domyślnych ustawień sterownika.
2. W przypadku tej nazwy DSN upewnij się, że wybrana jest opcja **Ze zintegrowanym uwierzytelnianiem systemu Windows**.
 - Jeśli produkt IBM SPSS Modeler i serwer IBM SPSS Modeler Server działają na różnych hostach, utwórz tę samą nazwę DSN ODBC na każdym hoście. Upewnij się, że ta sama nazwa DSN jest używana na każdym hoście.

Aktywacja integracji z usługami Analysis Services w produkcie IBM SPSS Modeler

Aby umożliwić produktowi IBM SPSS Modeler korzystanie z usług Analysis Services, należy najpierw udostępnić specyfikację serwera w oknie dialogowym Aplikacje pomocnicze.

1. Z menu programu IBM SPSS Modeler wybierz:
Narzędzia > Opcje > Aplikacje pomocnicze
2. Kliknij kartę **Microsoft**.
 - **Włącz integrację z Microsoft Analysis Services.** Włącza paletę Modelowanie w bazie (jeśli nie jest jeszcze wyświetlana) u dołu okna IBM SPSS Modeler i dodaje węzły przeznaczone dla algorytmów Analysis Services.
 - **Serwer hosta analiz.** Określ nazwę komputera, na którym uruchomione są usługi Analysis Services.
 - **Baza danych Analysis Server.** Wybierz żadaną bazę danych, klikając przycisk z wielokropkiem (...), aby otworzyć podrzędne okno dialogowe, w którym można wybierać spośród dostępnych baz danych. Ta lista jest wypełniona bazami danych dostępnymi dla określonego serwera analiz. Usługi Microsoft Analysis Services zapisują modele eksploracji danych w nazwanych bazach danych, dlatego należy wybrać odpowiednią bazę danych, w której zapisywane są modele Microsoft zbudowane przez produkt IBM SPSS Modeler.

- **Połączenie z serwerem SQL.** Określ nazwę DSN używaną przez bazę danych SQL Server w celu zapisywania danych, które są przekazywane do serwera analiz. Wybierz źródło danych ODBC, które będzie używane do udostępniania danych w celu budowania modeli eksploracji danych w usługach Analysis Services. Jeśli budujesz modele usług Analysis Services z danych dostarczanych w plikach płaskich lub źródłach danych ODBC, wówczas dane będą automatycznie wczytywane do tabeli tymczasowej utworzonej w bazie danych SQL Server, na którą wskazuje to źródło danych ODBC.
- **Wyświetl ostrzeżenie o możliwym zastąpieniu modelu.** Wybierz tę opcję, aby upewnić się, że modele zapisane w bazie danych nie będą zastępowane przez produkt IBM SPSS Modeler bez ostrzeżenia.

Uwaga: ustawienia wybrane w oknie dialogowym Aplikacje pomocnicze mogą być przesłaniane w różnych węzłach usług Analysis Services.

Włączenie opcji generowania i optymalizacji kodu SQL

1. Z menu programu IBM SPSS Modeler wybierz:
Narzędzia > Właściwości strumienia > Opcje
2. Kliknij opcję **Optymalizacja** w panelu nawigacji.
3. Upewnij się, że włączona jest opcja **Generuj kod SQL kierowany do bazy**. To ustawienie jest niezbędne, ponieważ zapewnia poprawne działanie modelowania w bazie danych.
4. Wybierz opcje **Optymalizuj operacje generujące kod SQL** i **Optymalizuj inne wykonywane operacje** (nie jest to ściśle wymagane, ale zdecydowanie zalecane w celu poprawy wydajności).

Budowanie modeli za pomocą usług Analysis Services

W przypadku budowania modelu w usługach Analysis Services wymagane jest, aby uczący zbiór danych znajdował się w tabeli lub widoku w bazie danych SQL Server. Jeśli dane nie znajdują się w bazie danych SQL Server lub muszą być przetwarzane w produkcie IBM SPSS Modeler w ramach procesów przygotowywania danych, które mogą być wykonywane w SQL Server, wówczas dane są automatycznie wczytywane do tabeli tymczasowej w SQL Server przed budowaniem modelu.

Zarządzanie modelami usług Analysis Services

Budowanie modelu usług Analysis Services za pośrednictwem produktu IBM SPSS Modeler powoduje utworzenie modelu w produkcie IBM SPSS Modeler oraz utworzenie lub zastąpienie modelu w bazie danych SQL Server. Model w produkcie IBM SPSS Modeler odwołuje się do zawartości modelu bazy danych zapisanego na serwerze bazy danych. Produkt IBM SPSS Modeler może przeprowadzić sprawdzanie spójności poprzez zapisanie identycznego wygenerowanego klucza tekstowego zarówno w modelu IBM SPSS Modeler, jak i w modelu SQL Server.



Węzeł modelowania **MS Decision Tree** jest stosowany w modelowaniu predykcyjnym atrybutów jakościowych i ilościowych. W przypadku atrybutów jakościowych ten węzeł generuje predykcje na podstawie relacji między kolumnami wejściowymi w zestawie danych. Przykładem niech będzie scenariusz przeznaczony do przewidzenia tego, którzy klienci kupią rower: jeśli dziewięciu z dziesięciu młodszych klientów kupi rower, ale zrobi to tylko dwóch z dziesięciu starszych klientów, wówczas węzeł wysnuje wniosek, że wiek jest dobrym predyktorem zakupu roweru. Drzewo decyzyjne generuje predykcje na podstawie tej tendencji w stronę konkretnego wyniku. W przypadku atrybutów ilościowych algorytm wykorzystuje regresję liniową w celu ustalenia miejsca podziału drzewa decyzyjnego. Jeśli więcej niż jedna kolumna jest ustawiona jako przewidywalna lub jeśli dane wejściowe zawierają tabelę zagnieżdżoną ustawioną jako przewidywalna, wówczas ten węzeł buduje osobne drzewo decyzyjne dla każdej kolumny przewidywalnej.



Węzeł modelowania **MS Clustering** stosuje techniki iteracyjne w celu grupowania obserwacji w zestawie danych w klastry o podobnych cechach. Te grupowania są użyteczne w przypadku eksploracji danych, identyfikowania nieprawidłowości w danych, a także w przypadku tworzenia predykcji. Modele grupowania identyfikują relacje w zestawie danych, które mogłyby zostać pominięte w logicznej analizie w przypadku swobodnej obserwacji. Na przykład można by wywnioskować, że ludzie dojeżdżający do pracy na rowerach zwykle pracują niedaleko. Algorytm może jednak znaleźć inne cechy dotyczące osób dojeżdżających na rowerach, które nie są oczywiste. Węzeł grupowania różni się od innych węzłów eksploracji danych, ponieważ nie jest określona żadna zmienna przewidywana. Węzeł grupowania uczy model wykorzystując wyłącznie relacje, jakie istnieją w danych, a także grupy, które identyfikuje węzeł.



Węzeł modelowania **MS Association Rules** jest użyteczny w przypadku mechanizmów rekomendacji. Mechanizm rekomendacji rekomenduje klientom produkty na podstawie pozycji, które już kupili lub które ich zainteresowały. Modele asocjacyjne są budowane na podstawie zestawów danych zawierających identyfikatory dla pojedynczych obserwacji oraz dla pozycji (elementów) zawartych w tych obserwacjach. Grupa pozycji w obserwacji jest nazywana **zbiorem elementów**. Model asocjacyjny jest złożony z serii zbiorów elementów oraz z reguł opisujących sposoby grupowania tych elementów w obserwacjach. Reguły, które algorytm identyfikuje, mogą być wykorzystywane do przewidywania prawdopodobnych zakupów klienta w przyszłości na podstawie elementów istniejących już w koszyku klienta.



Węzeł modelowania **MS Naive Bayes** oblicza prawdopodobieństwo warunkowe między zmienną przewidywaną a predyktorem i zakłada, że kolumny są niezależne. Model jest określany jako naiwny, ponieważ traktuje wszystkie proponowane zmienne predykcji jako niezależne od siebie. Ta metoda jest mniej wymagająca obliczeniowo niż inne algorytmy usługi Analysis Services i dlatego jest bardziej użyteczna w przypadku wykrywania relacji w początkowych etapach modelowania. Za pomocą tego węzła można przeprowadzać wstępne eksploracje danych, a następnie stosować wyniki w celu tworzenia dodatkowych modeli z innymi węzłami, których obliczanie może trwać dłużej, ale zwracane są bardziej dokładne wyniki.



Węzeł modelowania **MS Linear Regression** jest odmianą węzła Drzewa decyzyjne, w którym parametr `MINIMUM_LEAF_CASES` jest ustawiony na wartość równą łącznej liczbie obserwacji w zestawie danych (lub większą od niej), z którego węzeł korzysta w celu uczenia modelu eksploracji. Gdy zestaw parametrów jest ustawiony w ten sposób, węzeł nigdy nie utworzy podziału i dlatego przeprowadza regresję liniową.



Węzeł modelowania **MS Neural Network** jest podobny do węzła MS Decision Tree, ponieważ węzeł MS Neural Network oblicza prawdopodobieństwa każdego możliwego stanu atrybutu wejściowego przy założeniu każdego stanu atrybutu przewidywanego. Te prawdopodobieństwa można następnie wykorzystać w celu przewidzenia wyniku atrybutu przewidywanego na podstawie atrybutów wejściowych.



Węzeł modelowania **MS Logistic Regression** jest odmianą węzła MS Neural Network, w którym parametr `HIDDEN_NODE_RATIO` jest ustawiony na 0. To ustawienie powoduje utworzenie sieci neuronowej, która nie zawiera ukrytej warstwy i dlatego jest równoważna regresji logistycznej.



Węzeł modelowania **MS Time Series** udostępnia algorytmy regresji zoptymalizowane pod kątem prognozowania w czasie wartości ciągłych, takich jak sprzedaż produktu. Inne algorytmy firmy Microsoft, takie jak drzewa decyzyjne, wymagają dodatkowych kolumn zawierających nowe informacje jako dane wejściowe wymagane do przewidzenia trendu, jednak model szeregów czasowych (Time Series) ich nie wymaga. Model szeregów czasowych może przewidywać trendy na podstawie oryginalnego zestawu danych używanego do utworzenia modelu. Ponadto w przypadku predykcji, gdy wymagane jest automatyczne uwzględnienie nowych danych w analizie trendu, możliwe jest dodawanie nowych danych do modelu. Więcej informacji można znaleźć w temacie “Węzeł MS Time Series” na stronie 19.



Węzeł modelowania **MS Sequence Clustering** rozpoznaje uporządkowane sekwencje w danych i łączy taką analizę z technikami grupowania, aby generować grupy na podstawie kolejności i innych atrybutów. Więcej informacji można znaleźć w temacie “Węzeł MS Sequence Clustering” na stronie 20.

Dostęp do każdego węzła można uzyskać z palety Modelowanie w bazie, która jest dostępna u dołu okna produktu IBM SPSS Modeler.

Ustawienia wspólne dla wszystkich węzłów algorytmów

Następujące ustawienia są wspólne dla wszystkich algorytmów usług Analysis Services.

Opcje serwera

Na karcie Serwer można skonfigurować hosta serwera i bazę danych analiz, a także źródło danych programu SQL Server. Opcje określone w tym miejscu zastępują określone na karcie Microsoft w oknie dialogowym Aplikacje pomocnicze. Więcej informacji można znaleźć w temacie “Aktywacja integracji z usługami Analysis Services” na stronie 13.

Uwaga: wariant tej karty jest również dostępny podczas oceniania modeli Analysis Services. Więcej informacji można znaleźć w temacie “Karta serwera modelu użytkowego usług Analysis Services” na stronie 21.

Opcje modelu

W celu zbudowania najbardziej podstawowego modelu należy określić opcje na karcie Model przed wykonaniem dalszych czynności. Opcja oceny i inne opcje zaawansowane są dostępne na karcie Zaawansowane.

Dostępne są następujące podstawowe opcje modelowania:

Nazwa modelu. Określa nazwę nadawaną modelowi utworzonemu z chwilą wykonania danego węzła.

- **Auto.** Nazwa modelu jest generowana automatycznie, na podstawie nazw zmiennej przewidywanej lub identyfikacyjnej lub na podstawie nazwy typu modelu — w przypadkach gdy nie jest określona zmienna przewidywana (na przykład w przypadku modeli skupień).
- **Użytkownika.** Umożliwia podanie nazwy niestandardowej dla utworzonego modelu.

Użyj danych podzielonych na podzbiory. Dzieli dane na osobne podzbiory lub próby na potrzeby uczenia, testowania i walidacji na podstawie aktualnej zmiennej dzielącej na podzbiory. Korzystając z jednej próby do utworzenia modelu oraz innej do testowania go, można uzyskać informacje na temat tego, jak dobrze model pozwala uogólnić większe zestawy danych, które są podobne do bieżących danych. Jeśli w strumieniu nie określono zmiennej dzielącej na podzbiory, wówczas ta opcja jest ignorowana.

Z pogłębianiem. Jeśli ta opcja jest widoczna, wówczas umożliwia generowanie zapytań dotyczących modelu w celu uzyskania szczegółowych informacji na temat obserwacji uwzględnionych w modelu.

Zmienna unikalna. Z listy rozwijanej wybierz zmienną, która jednoznacznie identyfikuje każdą obserwację. Zwykle jest to zmienna identyfikacyjna, taka jak **CustomerID**.

Opcje zaawansowane MS Decision Tree

Opcje dostępne na karcie Zaawansowane mogą ulegać zmianie w zależności od struktury wybranego strumienia. Pełne szczegółowe informacje dotyczące opcji zaawansowanych dla wybranego węzła modelu Analysis Services są dostępne w pierwszym poziomie pomocy w interfejsie użytkownika.

Opcje zaawansowane MS Clustering

Opcje dostępne na karcie Zaawansowane mogą ulegać zmianie w zależności od struktury wybranego strumienia. Pełne szczegółowe informacje dotyczące opcji zaawansowanych dla wybranego węzła modelu Analysis Services są dostępne w pierwszym poziomie pomocy w interfejsie użytkownika.

Opcje zaawansowane MS Naive Bayes

Opcje dostępne na karcie Zaawansowane mogą ulegać zmianie w zależności od struktury wybranego strumienia. Pełne szczegółowe informacje dotyczące opcji zaawansowanych dla wybranego węzła modelu Analysis Services są dostępne w pierwszym poziomie pomocy w interfejsie użytkownika.

Opcje zaawansowane MS Linear Regression

Opcje dostępne na karcie Zaawansowane mogą ulegać zmianie w zależności od struktury wybranego strumienia. Pełne szczegółowe informacje dotyczące opcji zaawansowanych dla wybranego węzła modelu Analysis Services są dostępne w pierwszym poziomie pomocy w interfejsie użytkownika.

Opcje zaawansowane MS Neural Network

Opcje dostępne na karcie Zaawansowane mogą ulegać zmianie w zależności od struktury wybranego strumienia. Pełne szczegółowe informacje dotyczące opcji zaawansowanych dla wybranego węzła modelu Analysis Services są dostępne w pierwszym poziomie pomocy w interfejsie użytkownika.

Opcje zaawansowane MS Logistic Regression

Opcje dostępne na karcie Zaawansowane mogą ulegać zmianie w zależności od struktury wybranego strumienia. Pełne szczegółowe informacje dotyczące opcji zaawansowanych dla wybranego węzła modelu Analysis Services są dostępne w pierwszym poziomie pomocy w interfejsie użytkownika.

Węzeł reguł asocjacyjnych MS

Węzeł modelowania MS Association Rules jest użyteczny w przypadku mechanizmów rekomendacji. Mechanizm rekomendacji rekomenduje klientom produkty na podstawie pozycji, które już kupili lub które ich zainteresowały. Modele asocjacyjne są budowane na podstawie zestawów danych zawierających identyfikatory dla pojedynczych obserwacji oraz dla pozycji (elementów) zawartych w tych obserwacjach. Grupa pozycji w obserwacji jest nazywana **zbiorem elementów**.

Model asocjacyjny jest złożony z serii zbiorów elementów oraz z reguł opisujących sposoby grupowania tych elementów w obserwacjach. Reguły, które algorytm identyfikuje, mogą być wykorzystywane do przewidywania prawdopodobnych zakupów klienta w przyszłości na podstawie elementów istniejących już w koszyku klienta.

W przypadku danych w formacie tabelarycznym ten algorytm tworzy oceny, które reprezentują prawdopodobieństwo (\$MP-zmienna) dla każdej wygenerowanej rekomendacji (\$M-zmienna). W przypadku danych w formacie transakcyjnym oceny są tworzone na potrzeby obsługi (\$MS-zmienna), prawdopodobieństwa (\$MP-zmienna) i prawdopodobieństwa skorygowanego (\$MAP-zmienna) dla każdej wygenerowanej rekomendacji (\$M-zmienna).

Wymagania

Wymagania dotyczące transakcyjnego modelu asocjacyjnego są następujące:

- **Zmienna unikalna.** Model reguł asocjacyjnych wymaga klucza, który jednoznacznie identyfikuje rekordy.
- **Zmienna identyfikacyjna.** Gdy model MS Association Rules jest budowany z użyciem danych w formacie transakcyjnym, wymagana jest zmienna identyfikacyjna, która identyfikuje każdą transakcję. Zmienne identyfikacyjne można ustawiać na tę samą wartość, co zmienna unikalna.
- **Wymagana jest co najmniej jedna zmienna wejściowa.** Algorytm Association Rules wymaga co najmniej jednej zmiennej wejściowej.
- **Zmienna przewidywana.** Podczas budowania modelu MS Association z danymi transakcyjnymi zmienna przewidywana musi być zmienną transakcyjną, na przykład produkty kupione przez użytkownika.

Opcje zaawansowane MS Association Rules

Opcje dostępne na karcie Zaawansowane mogą ulegać zmianie w zależności od struktury wybranego strumienia. Pełne szczegółowe informacje dotyczące opcji zaawansowanych dla wybranego węzła modelu Analysis Services są dostępne w pierwszym poziomie pomocy w interfejsie użytkownika.

Węzeł MS Time Series

Węzeł modelowania MS Time Series obsługuje dwa rodzaje predykcji:

- przyszłe
- historyczne

Predykcje przyszłe oszacowują wartości zmiennych przewidywanych przez określoną liczbę okresów po zakończeniu danych historycznych użytkownika i są wykonywane zawsze. **Predykcje historyczne** to wartości zmiennych przewidywanych oszacowane przez określoną liczbę okresów, dla których użytkownik posiada rzeczywiste wartości w danych historycznych. Predykcje historyczne można wykorzystać w celu oceny jakości modelu poprzez porównanie rzeczywistych wartości historycznych z wartościami przewidywanymi. Wartość punktu początkowego dla predykcji określa, czy wykonywane są predykcje historyczne.

Węzeł MS Time Series — w przeciwieństwie do węzła Szeregi czasowe produktu IBM SPSS Modeler — nie potrzebuje poprzedzającego węzła Przedziały czasowe. Kolejna różnica polega na tym, że domyślnie oceny są generowane tylko dla wierszy przewidywanych, a nie dla wierszy historycznych w danych szeregu czasowego.

Wymagania

Wymagania dotyczące modelu MS Time Series są następujące:

- **Pojedyncza kluczowa zmienna czasu.** Każdy model musi zawierać jedną zmienną liczbową lub zmienną daty, która będzie używana jako szereg obserwacji definiujący przedziały czasu, których będzie używać model. Kluczowa zmienna czasu może być typu Data/czas lub typu liczbowego. Jednak zmienna musi zawierać wartości ciągłe i wartości muszą być unikalne dla każdego szeregu.
- **Pojedyncza zmienna przewidywana.** W każdym modelu można określić tylko jedną zmienną przewidywaną. Typ danych zmiennej przewidywanej musi obejmować wartości ciągłe. Na przykład można przewidzieć sposób zmiany atrybutów liczbowych, takich jak dochód, sprzedaż lub temperatura. Jednak w postaci zmiennej przewidywanej nie można użyć zmiennej, która zawiera wartości jakościowe, np. status zakupu lub wykształcenie.
- **Wymagana jest co najmniej jedna zmienna wejściowa.** Algorytm MS Time Series wymaga co najmniej jednej zmiennej wejściowej. Typ danych zmiennej wejściowej musi obejmować wartości ciągłe. Zmienne wejściowe nieciągłe są ignorowane podczas budowania modelu.
- **Zbiór danych musi być sortowany.** Wejściowy zbiór danych musi być sortowany (według kluczowej zmiennej czasu). W przeciwnym wypadku budowanie modelu zostanie przerwane z błędem.

Opcje modelu MS Time Series

Nazwa modelu. Określa nazwę nadawaną modelowi utworzonemu z chwilą wykonania danego węzła.

- **Auto.** Nazwa modelu jest generowana automatycznie, na podstawie nazw zmiennej przewidywanej lub identyfikacyjnej lub na podstawie nazwy typu modelu — w przypadkach gdy nie jest określona zmienna przewidywana (na przykład w przypadku modeli skupień).
- **Użytkownika.** Umożliwia podanie nazwy niestandardowej dla utworzonego modelu.

Użyj danych podzielonych na podzbiory. Jeśli zdefiniowano zmienną dzielącą na podzbiory, ta opcja umożliwia użycie podczas budowania modelu wyłącznie danych z podzbioru uczącego.

Z pogłębianiem. Jeśli ta opcja jest widoczna, wówczas umożliwia generowanie zapytań dotyczących modelu w celu uzyskania szczegółowych informacji na temat obserwacji uwzględnionych w modelu.

Zmienna unikalna. Z listy rozwijanej wybierz kluczową zmienną czasu, która jest używana do budowania modelu szeregu czasowego.

Opcje zaawansowane MS Time Series

Opcje dostępne na karcie Zaawansowane mogą ulegać zmianie w zależności od struktury wybranego strumienia. Pełne szczegółowe informacje dotyczące opcji zaawansowanych dla wybranego węzła modelu Analysis Services są dostępne w pierwszym poziomie pomocy w interfejsie użytkownika.

Jeśli wykonujesz predykcje historyczne, liczba etapów historycznych, które mogą zostać uwzględnione w wyniku oceny, jest określana przez wartość ($HISTORIC_MODEL_COUNT * HISTORIC_MODEL_GAP$). Domyślnie ograniczenie ma wartość 10, co oznacza, że zostanie wykonanych tylko 10 predykcji historycznych. W takim przypadku (przykładowym) wystąpi błąd w przypadku wprowadzenia wartości mniejszej niż -10 dla opcji **Predykcja historyczna** na karcie Ustawienia w modelu użytkowym (patrz “Karta ustawienia modelu użytkowego MS Time Series” na stronie 23). W celu uzyskania większej liczby predykcji historycznych można zwiększyć wartość $HISTORIC_MODEL_COUNT$ lub $HISTORIC_MODEL_GAP$, ale spowoduje to zwiększenie czasu budowania modelu.

Opcje ustawień MS Time Series

Początek okresu oszacowania. Określ okres, w którym rozpoczną się predykcje.

- **Rozpocznij od: Nowa predykcja.** Okres, w którym mają się rozpocząć przyszłe predykcje, wyrażony jako przesunięcie od ostatniego okresu w danych historycznych użytkownika. Na przykład, jeśli dane historyczne zakończyły się 12/99, a chcesz rozpocząć predykcje od 01/00, użyj wartości 1; jeśli jednak predykcje mają się rozpocząć od 03/00, użyj wartości 3.
- **Rozpocznij od: Predykcja historyczna.** Okres, w którym mają się rozpocząć predykcje historyczne, wyrażony jako przesunięcie ujemne od ostatniego okresu w danych historycznych użytkownika. Na przykład, jeśli dane historyczne zakończyły się 12/99, a chcesz uzyskać predykcje historyczne dla ostatnich pięciu okresów Twoich danych, użyj wartości -5.

Koniec okresu oszacowania. Określ okres, w którym predykcje zostaną zatrzymane.

- **Końcowy krok predykcji.** Okres, w którym predykcje zostaną zatrzymane, wyrażony jako przesunięcie od ostatniego okresu w danych historycznych użytkownika. Na przykład, jeśli dane historyczne zakończyły się 12/99, a chcesz, aby predykcje zatrzymały się 6/00, użyj w tej opcji wartości 6. W przypadku przyszłych predykcji wartość musi być zawsze większa od lub równa wartości **Rozpocznij od**.

Węzeł MS Sequence Clustering

Węzeł MS Sequence Clustering używa algorytmu analizy sekwencyjnej, który eksploruje dane zawierające zdarzenia, które mogą zostać powiązane poprzez śledzenie ścieżek lub *sekwencji*. Do przykładów można zaliczyć ścieżki klikania utworzone, gdy użytkownicy nawigują lub przeglądają stronę WWW, a także kolejność, w jakiej klient dodaje artykuły do koszyka w sklepie online. Algorytm znajduje najbardziej typowe sekwencje na podstawie *grupowania* sekwencji, które są identyczne.

Wymagania

Wymagania dotyczące modelu Microsoft Sequence Clustering są następujące:

- **Zmienna identyfikacyjna.** Algorytm Microsoft Sequence Clustering wymaga, aby informacje o sekwencji były zapisane w formacie transakcyjnym. W celu zrealizowania tego wymagania konieczna jest zmienna identyfikacyjna, która będzie identyfikować każdą transakcję.
- **Wymagana jest co najmniej jedna zmienna wejściowa.** Algorytm wymaga co najmniej jednej zmiennej wejściowej.
- **Zmienna sekwencji.** Algorytm wymaga także zmiennej identyfikacyjnej sekwencji, która musi mieć poziom pomiaru ciągły. Na przykład można użyć identyfikatora strony WWW, wartości całkowitej lub ciągu znaków, pod warunkiem że zmienna identyfikuje zdarzenia w sekwencji. Dla każdej sekwencji dozwolony jest jeden identyfikator i tylko jeden typ sekwencji jest dozwolony dla każdego modelu. Zmienna sekwencji musi się różnić od zmiennych identyfikacyjnych i zmiennych unikalnych.
- **Zmienna przewidywana.** Zmienna przewidywana jest wymagana podczas budowania modelu grupowania sekwencyjnego.
- **Zmienna unikalna.** Model grupowania sekwencyjnego wymaga zmiennej kluczowej, która jednoznacznie identyfikuje rekordy. Zmienną unikalną można ustawić w taki sam sposób, jak zmienną identyfikacyjną.

Opcje zmiennej MS Sequence Clustering

Wszystkie węzły modelowania zawierają kartę Zmienne, na której można określić zmienne, które będą używane podczas budowania modelu.

Aby możliwe było zbudowanie modelu grupowania sekwencyjnego, konieczne jest określenie, które zmienne mają być używane jako zmienne przewidywane, a które jako dane wejściowe. Należy zwrócić uwagę na to, że w przypadku węzła MS Sequence Clustering nie można używać informacji o zmiennej z poprzedzającego węzła Typ; ustawienia zmiennej należy określić w tym miejscu.

Identyfikator. Należy wybrać zmienną identyfikacyjną z listy. Jako zmienna identyfikacyjna mogą być używane zmienne numeryczne lub symboliczne. Każda unikalna wartość tej zmiennej powinna wskazywać na określoną jednostkę analizy. Na przykład w aplikacji do obsługi koszyka zakupów każdy identyfikator może reprezentować jednego klienta. W przypadku aplikacji do analizy dzienników sieciowych każdy identyfikator może reprezentować komputer (wg adresu IP) lub użytkownika (wg danych logowania).

Zmienne wejściowe. Wybierz zmienną wejściową (lub zmienne) dla modelu. Są to zmienne zawierające zdarzenia interesujące w modelowaniu sekwencji.

Sekwencja. Wybierz zmienną z listy, która będzie używana jako zmienna identyfikacyjna sekwencji. Na przykład można użyć identyfikatora strony WWW, wartości całkowitej lub ciągu znaków, pod warunkiem że zmienna identyfikuje zdarzenia w sekwencji. Dla każdej sekwencji dozwolony jest jeden identyfikator i tylko jeden typ sekwencji jest dozwolony dla każdego modelu. Zmienna sekwencji musi się różnić od zmiennej identyfikacyjnej (określonej na tej karcie), a także od zmiennej unikalnej (określonej na karcie Model).

Zmienna przewidywana. Wybierz zmienną, która będzie używana jako zmienna przewidywana, czyli zmienna, której wartość próbujesz przewidzieć na podstawie danych sekwencji.

Opcje zaawansowane MS Sequence Clustering

Opcje dostępne na karcie Zaawansowane mogą ulegać zmianie w zależności od struktury wybranego strumienia. Pełne szczegółowe informacje dotyczące opcji zaawansowanych dla wybranego węzła modelu Analysis Services są dostępne w pierwszym poziomie pomocy w interfejsie użytkownika.

Ocenianie modelami usług Analysis Services

Ocenianie modelu odbywa się w programie SQL Server i jest wykonywane przez usługi Analysis Services. Jeśli dane pochodzą z produktu IBM SPSS Modeler lub muszą być w nim przygotowywane w IBM SPSS Modeler, wówczas konieczne może być wczytanie zestawu danych do tabeli tymczasowej. Modele tworzone z produktu IBM SPSS Modeler przy użyciu eksploracji w bazie danych stanowią w rzeczywistości model zdalny przechowywany w zdalnym serwerze eksploracji danych lub serwerze bazy danych. Jest to istotna różnica, którą należy zrozumieć podczas przeglądania i oceny modeli utworzonych przy użyciu algorytmów usług Microsoft Analysis Services.

W produkcie IBM SPSS Modeler udostępniana jest zwykle tylko jedna predykcja i powiązane z nią prawdopodobieństwo lub współczynnik ufności.

Przykłady oceniania modeli przedstawiono w sekcji “Przykłady eksploracji w usługach Analysis Services” na stronie 24.

Ustawienia wspólne dla wszystkich algorytmów usług Analysis Services

Następujące ustawienia są wspólne dla wszystkich modeli usług Analysis Services.

Karta serwera modelu użytkowego usług Analysis Services

Karta Serwer służy do określania połączeń na potrzeby eksploracji w bazie danych. Ta karta udostępnia także unikatowy klucz modelu. Klucz jest losowo generowany podczas budowania modelu i zapisywany w modelu w produkcie IBM SPSS Modeler, a także w opisie obiektu modelu w bazie danych usług Analysis Services.

Na karcie Serwer można skonfigurować hosta serwera i bazę danych analiz, a także źródło danych SQL Server na potrzeby operacji oceniania. Opcje określone w tym miejscu zastępują określone w oknie dialogowym Aplikacje

pomocnicze i w oknie budowy modelu w produkcie IBM SPSS Modeler. Więcej informacji można znaleźć w temacie “Aktywacja integracji z usługami Analysis Services” na stronie 13.

Identyfikator GUID modelu. W tym miejscu widoczny jest klucz modelu. Klucz jest losowo generowany podczas budowania modelu i zapisywany w modelu w produkcie IBM SPSS Modeler, a także w opisie obiektu modelu w bazie danych usług Analysis Services.

Sprawdź. Kliknięcie tego przycisku umożliwi sprawdzenie klucza modelu poprzez porównanie go z kluczem w modelu zapisanym w bazie danych usług Analysis Services. Dzięki temu można sprawdzić, czy model nadal istnieje na serwerze analizy i wskazuje, że struktura modelu nie uległa zmianie.

Uwaga: przycisk Sprawdź jest dostępny tylko dla modeli dodanych do obszaru roboczego strumienia podczas przygotowań do oceny. Jeśli sprawdzenie nie powiedzie się, należy ustalić, czy model został usunięty lub zastąpiony innym modelem na serwerze.

Widok. Kliknij, aby uzyskać widok graficzny modelu drzewa decyzyjnego. Przeglądarka drzewa decyzyjnego jest współużytkowana przez inne algorytmy drzewa decyzyjnego w produkcie IBM SPSS Modeler, a jej sposób działania jest identyczny.

Karta podsumowania modelu użytkowego usług Analysis Services

Karta Podsumowanie modelu użytkowego zawiera informacje na temat samego modelu (*Analiza*), zmiennych użytych w modelu (*Zmienne*), ustawień użytych podczas budowania modelu (*Ustawienia budowania*) i uczenia modelu (*Podsumowanie uczenia*).

Podczas przeglądania węzła po raz pierwszy karta Podsumowanie jest zwinięta. Aby zobaczyć wyniki będące przedmiotem zainteresowania, należy użyć rozszerzanego elementu sterującego po lewej stronie pozycji, aby ją rozwinąć, lub kliknąć przycisk **Rozwiń wszystko**, aby wyświetlić wszystkie wyniki. W celu ukrycia wyników po zakończeniu ich przeglądania należy użyć rozszerzanego elementu sterującego, aby zwinąć konkretne wyniki, jakie mają zostać ukryte, lub kliknąć przycisk **Zwiń wszystko**, aby zwinąć wszystkie wyniki.

Analiza. Wyświetla informacje na temat konkretnego modelu. Jeśli wykonano węzeł analizy dołączony do modelu użytkowego, wówczas informacje z tej analizy również pojawiają się w tej sekcji.

Zmienne. Na liście znajdują się zmienne użyte jako zmienne przewidywane i wejściowe podczas budowania modelu.

Ustawienia budowania. Zawiera informacje na temat ustawień użytych podczas budowania modelu.

Podsumowanie uczenia. Przedstawia typ modelu, strumień użyty do jego utworzenia, użytkownika, który go utworzył, informację, kiedy został utworzony, oraz czas, jaki był potrzebny do zbudowania modelu.

Model użytkowy MS Time Series

Model MS Time Series generuje oceny tylko dla przewidywanych okresów czasowych, a nie dla danych historycznych.

W poniższej tabeli przedstawiono zmienne dodawane do modelu.

Tabela 1. Zmienne dodawane do modelu

Nazwa zmiennej	Opis
\$M-zmienna	Wartość przewidywana zmiennej
\$Var-zmienna	Wyliczona wariancja zmiennej
\$Stdev-zmienna	Odchylenie standardowe zmiennej

Karta serwera modelu użytkowego MS Time Series

Karta Serwer służy do określania połączeń na potrzeby eksploracji w bazie danych. Ta karta udostępnia także unikatowy klucz modelu. Klucz jest losowo generowany podczas budowania modelu i zapisywany w modelu w produkcie IBM SPSS Modeler, a także w opisie obiektu modelu w bazie danych usług Analysis Services.

Na karcie Serwer można skonfigurować hosta serwera i bazę danych analiz, a także źródło danych SQL Server na potrzeby operacji oceniania. Opcje określone w tym miejscu zastępują określone w oknie dialogowym Aplikacje pomocnicze i w oknie budowy modelu w produkcie IBM SPSS Modeler. Więcej informacji można znaleźć w temacie “Aktywacja integracji z usługami Analysis Services” na stronie 13.

Identyfikator GUID modelu. W tym miejscu widoczny jest klucz modelu. Klucz jest losowo generowany podczas budowania modelu i zapisywany w modelu w produkcie IBM SPSS Modeler, a także w opisie obiektu modelu w bazie danych usług Analysis Services.

Sprawdź. Kliknięcie tego przycisku umożliwi sprawdzenie klucza modelu poprzez porównanie go z kluczem w modelu zapisanym w bazie danych usług Analysis Services. Dzięki temu można sprawdzić, czy model nadal istnieje na serwerze analizy i wskazuje, że struktura modelu nie uległa zmianie.

Uwaga: przycisk Sprawdz jest dostępny tylko dla modeli dodanych do obszaru roboczego strumienia podczas przygotowań do oceny. Jeśli sprawdzenie nie powiedzie się, należy ustalić, czy model został usunięty lub zastąpiony innym modelem na serwerze.

Widok. Kliknij, aby uzyskać widok graficzny modelu szeregu czasowego. W usługach Analysis Services ukończony model zostanie wyświetlony jako drzewo. Zawsze można wyświetlić wykres przedstawiający wartość historyczną zmiennej przewidywanej w czasie, a także przewidywane wartości przyszłe.

Więcej informacji zawiera opis przeglądarki szeregów czasowych w bibliotece MSDN na stronie <http://msdn.microsoft.com/en-us/library/ms175331.aspx>.

Karta ustawienia modelu użytkowego MS Time Series

Początek okresu oszacowania. Określ okres, w którym rozpoczną się predykcje.

- **Rozpocznij od: Nowa predykcja.** Okres, w którym mają się rozpocząć przyszłe predykcje, wyrażony jako przesunięcie od ostatniego okresu w danych historycznych użytkownika. Na przykład, jeśli dane historyczne zakończyły się 12/99, a chcesz rozpocząć predykcje od 01/00, użyj wartości 1; jeśli jednak predykcje mają się rozpocząć od 03/00, użyj wartości 3.
- **Rozpocznij od: Predykcja historyczna.** Okres, w którym mają się rozpocząć predykcje historyczne, wyrażony jako przesunięcie ujemne od ostatniego okresu w danych historycznych użytkownika. Na przykład, jeśli dane historyczne zakończyły się 12/99, a chcesz uzyskać predykcje historyczne dla ostatnich pięciu okresów Twoich danych, użyj wartości -5.

Koniec okresu oszacowania. Określ okres, w którym predykcje zostaną zatrzymane.

- **Końcowy krok predykcji.** Okres, w którym predykcje zostaną zatrzymane, wyrażony jako przesunięcie od ostatniego okresu w danych historycznych użytkownika. Na przykład, jeśli dane historyczne zakończyły się 12/99, a chcesz, aby predykcje zatrzymały się 6/00, użyj w tej opcji wartości 6. W przypadku przyszłych predykcji wartość musi być zawsze większa od lub równa wartości **Rozpocznij od**.

Model użytkowy MS Sequence Clustering

W poniższej tabeli przedstawiono zmienne dodawane do modelu MS Sequence Clustering (gdzie *zmienna* jest nazwą zmiennej przewidywanej).

Tabela 2. Zmienne dodawane do modelu

Nazwa zmiennej	Opis
\$MC-zmienna	Predykcja skupienia, do którego należy sekwencja.

Tabela 2. Zmienne dodawane do modelu (kontynuacja)

Nazwa zmiennej	Opis
\$MCP-zmienna	Prawdopodobieństwo, że ta sekwencja należy do przewidywanego skupienia.
\$MS-zmienna	Wartość przewidywana <i>zmiennej</i>
\$MSP-zmienna	Prawdopodobieństwo, że wartość \$MS-zmienna jest poprawna.

Eksportowanie modeli i generowanie węzłów

Podsumowanie i strukturę modelu można wyeksportować do plików tekstowych i plików HTML. W razie potrzeby można generować odpowiednie węzły selekcji i filtrowania.

Model Microsoft Analysis Services, podobnie jak inne modele użytkowe w produkcie IBM SPSS Modeler, obsługuje bezpośrednie generowanie węzłów operacji na rekordach i zmiennych. Opcje dostępne w menu generowania w modelu użytkowym umożliwiają wygenerowanie następujących węzłów:

- Węzeł selekcji (pod warunkiem że element jest wybrany na karcie Model)
- węzeł Filtruj

Przykłady eksploracji w usługach Analysis Services

Poniżej przedstawiono szereg przykładowych strumieni, które demonstrują sposób użycia opcji eksploracji danych w usługach MS Analysis Services przy użyciu produktu IBM SPSS Modeler. Te strumienie można znaleźć w folderze instalacyjnym produktu IBM SPSS Modeler w katalogu:

`\Demos\Database_Modelling\Microsoft`

Uwaga: dostęp do folderu Demos można uzyskać z grupy programu IBM SPSS Modeler w menu Start systemu Windows.

Przykładowe strumienie: Drzewa decyzyjne

Następujące strumienie mogą być używane razem sekwencyjnie jako przykład procesu eksploracji bazy danych przy użyciu algorytmu Drzewa decyzyjne udostępnianego przez usługi MS Analysis Services.

Tabela 3. Drzewa decyzyjne — przykładowe strumienie

Strumień	Opis
<code>1_upload_data.str</code>	Służy do kasowania i wczytywania danych z pliku płaskiego do bazy danych.
<code>2_explore_data.str</code>	Udostępnia przykład eksploracji danych z użyciem produktu IBM SPSS Modeler
<code>3_build_model.str</code>	Buduje model z użyciem algorytmu rodzimego dla bazy danych.
<code>4_evaluate_model.str</code>	Używany jako przykład oceny modelu za pomocą produktu IBM SPSS Modeler
<code>5_deploy_model.str</code>	Wdraża model na potrzeby oceniania w bazie danych.

Uwaga: w celu uruchomienia tego przykładu strumienie muszą być wykonywane kolejno. Ponadto węzły źródłowe i węzły modelowania w każdym strumieniu muszą być aktualizowane w taki sposób, aby odwoływały się do poprawnego źródła danych dla bazy danych przeznaczonej do użycia.

Zestaw danych używany w strumieniach przykładowych dotyczy aplikacji kart kredytowych i prezentuje problem klasyfikacji z mieszaniną predyktorów jakościowych i ciągłych. Więcej informacji na temat tego zestawu danych zawiera plik `crx.names`, który znajduje się w tym samym folderze, w którym są przykładowe strumienie.

Ten zestaw danych jest dostępny z repozytorium UCI Machine Learning Repository na stronie <ftp://ftp.ics.uci.edu/pub/machine-learning-databases/credit-screening/>.

Przykładowy strumień: Wczytywanie danych

Pierwszy przykładowy strumień o nazwie *1_upload_data.str* jest używany do kasowania i wczytywania danych z pliku płaskiego do programu SQL Server.

Eksploracja danych za pomocą usług Analysis Services wymaga zmiennej kluczowej, dlatego ten strumień początkowo używa węzła wyliczeń, aby dodać nową zmienną do zestawu danych o nazwie *KEY* i unikalnych wartościach *1,2,3*, przy użyciu funkcji IBM SPSS Modeler *@INDEX*.

Następny w kolejności węzeł wypełniania jest używany do obsługi wartości brakujących i zastępuje puste pola wczytane z pliku tekstowego *ocrx.data* wartościami *NULL*.

Przykładowy strumień: Eksploracja danych

Drugi przykładowy strumień — *2_explore_data.str* — jest używany do zaprezentowania użycia węzła Audyt danych w celu uzyskania ogólnego przeglądu danych, w tym statystyk i wykresów podsumowujących.

Dwukrotne kliknięcie wykresu w raporcie z audytu danych generuje bardziej szczegółowy wykres, który pozwala na głębszą eksplorację pod kątem konkretnej zmiennej.

Przykładowy strumień: Budowa modelu

Trzeci przykładowy strumień — *3_build_model.str* — ilustruje budowanie modelu w produkcie IBM SPSS Modeler. Model bazy danych można dołączyć do strumienia, a następnie kliknąć dwukrotnie, aby określić ustawienia budowania.

Na karcie Model w oknie dialogowym można określić:

1. Wybierz zmienną **Klucz** jako zmienną o unikalnym identyfikatorze.

Na karcie Zaawansowany można precyzyjnie dostosowywać ustawienia dotyczące budowania modelu.

Przed uruchomieniem upewnij się, że określona została właściwa baza danych na potrzeby budowania modelu. W celu modyfikacji wszelkich ustawień użyj karty Serwer.

Przykładowy strumień: Ocena modelu

Czwarty przykładowy strumień — *4_evaluate_model.str* — ilustruje zalety korzystania z produktu IBM SPSS Modeler na potrzeby modelowania w bazie danych. Po wykonaniu modelu można go dodać z powrotem do strumienia danych i ocenić model, używając kilku narzędzi oferowanych w produkcie IBM SPSS Modeler.

Wyświetlanie wyników modelowania

Kliknij dwukrotnie model użytkowy, aby zapoznać się z wynikami. Na karcie Podsumowanie dostępny jest widok wyników z drzewa reguł. Można również kliknąć przycisk **Widok** dostępny na karcie Serwer, aby uzyskać widok graficzny modelu Drzewa decyzyjne.

Ocena wyników modelu

Węzeł analizy w strumieniu przykładowym tworzy macierz zbieżności przedstawiającą wzorzec dopasowań między poszczególnymi zmiennymi przewidywanymi a odpowiadającymi im zmiennymi zależnymi. Wykonaj węzeł analizy, aby wyświetlić wyniki.

Węzeł ewaluacji w strumieniu przykładowym może tworzyć wykres korzyści przeznaczony do przedstawienia postępów modelu w zakresie dokładności. Wykonaj węzeł ewaluacji, aby wyświetlić wyniki.

Przykładowy strumień: Wdrożenie modelu

Gdy dokładność modelu jest zadowalająca, można go wdrożyć, aby używać go z aplikacjami zewnętrznymi albo opublikować z powrotem do bazy danych. W ostatnim strumieniu przykładowym o nazwie `5_deploy_model.str` dane są odczytywane z tabeli CREDIT, a następnie oceniane i publikowane do tabeli CREDITSCORES przy użyciu węzła eksportu do bazy danych.

Uruchomienie tego strumienia powoduje wygenerowanie następującego kodu SQL:

```
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 [field15], 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]
)
T0
```

Rozdział 4. Modelowanie w bazie danych z użyciem rozwiązania Oracle Data Mining

Informacje o rozwiązaniu Oracle Data Mining

Produkt IBM SPSS Modeler umożliwia integrację z rozwiązaniem Oracle Data Mining (ODM), które udostępnia rodzinę algorytmów eksploracji danych osadzoną w systemie RDBMS Oracle. Dostęp do tych funkcji można uzyskać za pośrednictwem graficznego interfejsu użytkownika produktu IBM SPSS Modeler i środowiska programistycznego zorientowanego na przepływ pracy, dzięki którym klienci mogą korzystać z algorytmów eksploracji danych oferowanych przez ODM.

Produkt IBM SPSS Modeler obsługuje integrację następujących algorytmów z rozwiązania Oracle Data Mining:

- Naive Bayes
- Adaptive Bayes
- SVM (ang. Support Vector Machine)
- Uogólnione modele liniowe (GLM)*
- Drzewo decyzyjne
- O-Cluster
- K-średnie
- NMF (ang. Nonnegative Matrix Factorization)
- Apriori
- MDL (ang. Minimum Descriptor Length)
- AI (ang. Attribute Importance)

* Tylko 11g R1

Wymagania dotyczące integracji z systemem Oracle

Następujące warunki stanowią warunki wstępne do przeprowadzenia modelowania w bazie danych z użyciem produktu Oracle Data Mining. W celu upewnienia się, że te warunki są spełnione, konieczne mogą być konsultacje z administratorem bazy danych.

- Produkt IBM SPSS Modeler uruchamiany w trybie lokalnym lub względem instalacji IBM SPSS Modeler Server w systemie Windows lub UNIX.
- Oracle 10gR2 lub 11gR1 (baza danych w wersji 10.2 lub wyższej) z opcją Oracle Data Mining.

Uwaga: 10gR2 zapewnia obsługę wszystkich algorytmów modelowania w bazie danych z wyjątkiem algorytmu Uogólnione modele liniowe (on wymaga wersji 11gR1).

- Źródło danych ODBC w celu nawiązywania połączeń z Oracle zgodnie z opisem poniżej.

Uwaga: Modelowanie w bazie danych i optymalizacja SQL wymagają włączenia na komputerze z programem IBM SPSS Modeler możliwości połączenia z serwerem IBM SPSS Modeler Server. Po włączeniu tej opcji można uzyskać dostęp do algorytmów baz danych, wstawić SQL do kolejki bezpośrednio z programu IBM SPSS Modeler i uzyskać dostęp do programu IBM SPSS Modeler Server. W celu sprawdzenia bieżącego statusu licencji należy wybrać z menu programu IBM SPSS Modeler następujące opcje.

Pomoc > Informacje o programie > Dodatkowe szczegóły

Po włączeniu możliwości połączenia na karcie Status licencji widoczna jest opcja **Aktywacja serwera**.

Aktywacja integracji z produktem Oracle

Aby włączyć integrację produktu IBM SPSS Modeler z produktem Oracle Data Mining, należy skonfigurować produkt Oracle, utworzyć źródło ODBC, włączyć integrację w oknie dialogowym Aplikacje pomocnicze w produkcie IBM SPSS Modeler, a także włączyć generowanie i optymalizację kodu SQL.

Konfigurowanie produktu Oracle

Informacje dotyczące instalowania i konfigurowania produktu Oracle Data Mining zawiera dokumentacja produktu — przede wszystkim dokumentacja *Oracle Administrator's Guide*.

Tworzenie źródła ODBC dla produktu Oracle

Aby umożliwić połączenie produktu Oracle i produktu IBM SPSS Modeler, należy utworzyć systemową nazwę źródła danych (DSN) ODBC.

Przed utworzeniem DSN wymagana jest podstawowa znajomość źródeł danych i sterowników ODBC, a także obsługa bazy danych w produkcie IBM SPSS Modeler.

W przypadku uruchamiania w trybie rozproszonym w produkcie IBM SPSS Modeler Server należy utworzyć DSN na komputerze serwera. W przypadku uruchamiania w trybie lokalnym (klienta) należy utworzyć DSN na komputerze klienckim.

1. Zainstaluj sterowniki ODBC. Są dostępne na dysku instalacyjnym produktu IBM SPSS Data Access Pack dostarczonym z tym wydaniem. Uruchom plik *setup.exe*, aby uruchomić instalatora i wybrać wszystkie odpowiednie sterowniki. W celu zainstalowania sterowników postępuj zgodnie z instrukcjami wyświetlanymi na ekranie.
 - a. Utwórz DSN.

Uwaga: kolejność opcji w menu jest zależna od wersji systemu Windows.

 - **Windows XP.** Z menu Start wybierz opcję **Panel sterowania**. Kliknij dwukrotnie opcję **Narzędzia administracyjne**, a następnie kliknij dwukrotnie opcję **Źródła danych (ODBC)**.
 - **Windows Vista.** Z menu Start wybierz opcję **Panel sterowania**, a następnie **Konserwacja systemu**. Kliknij dwukrotnie pozycję **Narzędzia administracyjne**, wybierz pozycję **Źródła danych (ODBC)**, a następnie kliknij opcję **Otwórz**.
 - **Windows 7.** Z menu Start wybierz pozycję **Panel sterowania**, następnie **System i zabezpieczenia**, a następnie **Narzędzia administracyjne**. Wybierz pozycję **Źródła danych (ODBC)**, a następnie kliknij opcję **Otwórz**.
 - b. Kliknij kartę **Systemowe źródło danych DSN** i kliknij opcję **Dodaj**.
2. Wybierz sterownik **SPSS OEM 6.0 Oracle Wire Protocol**.
3. Kliknij przycisk **Zakończ**.
4. Na ekranie ODBC Oracle Wire Protocol Driver Setup wprowadź wybraną nazwę źródła danych, nazwę hosta serwera Oracle, numer portu dla połączenia, a także identyfikator SID dla instancji Oracle, której używasz.

Nazwa hosta, port i identyfikator SID można uzyskać z pliku *tnsnames.ora* na komputerze serwera, pod warunkiem że zaimplementowano TNS z plikiem *tnsnames.ora*. W celu uzyskania dodatkowych informacji skontaktuj się z administratorem produktu Oracle.
5. Kliknij przycisk **Test**, aby sprawdzić połączenie.

Włączanie integracji z produktem Oracle Data Mining w produkcie IBM SPSS Modeler

1. Z menu programu IBM SPSS Modeler wybierz:
Narzędzia > Opcje > Aplikacje pomocnicze
2. Kliknij kartę **Oracle**.

Włącz integrację Oracle Data Mining. Włącza paletę Modelowanie w bazie (jeśli nie jest jeszcze wyświetlana) u dołu okna IBM SPSS Modeler i dodaje węzły przeznaczone dla algorytmów Oracle Data Mining.

Połączenie z bazą Oracle. Określ domyślne źródło danych ODBC Oracle używane na potrzeby budowania i zapisywania modeli, a także poprawną nazwę użytkownika i hasło użytkownika. To ustawienie może zostać przesłonięte w pojedynczych węzłach modelowania i węzłach użytkowych.

Uwaga: połączenie z bazą danych używane na potrzeby modelowania może być takie samo, jak połączenie używane w celu uzyskiwania dostępu do danych. Na przykład może istnieć strumień, który uzyskuje dostęp do danych z jednej z baz danych Oracle, pobiera dane do produktu IBM SPSS Modeler w celu kasowania lub w celu przeprowadzenia innych manipulacji, a następnie przesyła te dane do innej bazy danych Oracle na potrzeby modelowania. Alternatywnie oryginalne dane mogą istnieć w pliku płaskim lub innym źródle (innym niż Oracle), ale w takim przypadku na potrzeby modelowania konieczne będzie przesłanie ich do produktu Oracle. We wszystkich przypadkach dane będą automatycznie wprowadzane do tabeli tymczasowej utworzonej w bazie danych używanej na potrzeby modelowania.

Ostrzegaj o możliwym nadpisaniu modelu Oracle Data Mining. Wybierz tę opcję, aby upewnić się, że modele zapisane w bazie danych nie będą zastępowane przez produkt IBM SPSS Modeler bez ostrzeżenia.

Wymień modele Oracle Data Mining. Ta opcja umożliwia wyświetlenie dostępnych modeli eksploracji danych.

Pozwól na uruchomienie Oracle Data Miner. (opcjonalnie) Gdy ta opcja jest włączona, produkt IBM SPSS Modeler może uruchamiać aplikację Oracle Data Miner. Więcej informacji zawiera sekcja “Oracle Data Miner” na stronie 46.

Ścieżka do programu Oracle Data Miner. (opcjonalnie) Ta opcja określa fizyczną lokalizację pliku wykonywalnego aplikacji Oracle Data Miner dla systemu Windows (na przykład `C:\odm\bin\odminerw.exe`). Aplikacja Oracle Data Miner nie jest instalowana za pomocą produktu IBM SPSS Modeler; właściwą wersję należy pobrać ze strony WWW Oracle (<http://www.oracle.com/technology/products/bi/odm/odminer.html>), a następnie zainstalować na kliencie.

Włączenie opcji generowania i optymalizacji kodu SQL

1. Z menu programu IBM SPSS Modeler wybierz:
Narzędzia > Właściwości strumienia > Opcje
2. Kliknij opcję **Optymalizacja** w panelu nawigacji.
3. Upewnij się, że włączona jest opcja **Generuj kod SQL kierowany do bazy**. To ustawienie jest niezbędne, ponieważ zapewnia poprawne działanie modelowania w bazie danych.
4. Wybierz opcje **Optymalizuj operacje generujące kod SQL** i **Optymalizuj inne wykonywane operacje** (nie jest to ściśle wymagane, ale zdecydowanie zalecane w celu poprawy wydajności).

Budowanie modeli z użyciem rozwiązania Oracle Data Mining

Węzły budowania modeli Oracle działają tak samo, jak inne węzły modelowania w produkcie IBM SPSS Modeler, lecz z kilkoma wyjątkami. Dostęp do tych węzłów można uzyskać z palety Modelowanie w bazie, która jest dostępna u dołu okna produktu IBM SPSS Modeler.

Zagadnienia dotyczące danych

Oracle wymaga, aby dane jakościowe były zapisywane w formacie łańcuchowym (CHAR lub VARCHAR2). W rezultacie produkt IBM SPSS Modeler nie zezwala na określanie zmiennych liczbowych o poziomie pomiaru *Flaga* ani *Nominalne* (jakościowe) jako danych wejściowych dla modeli ODM. W razie potrzeby liczby mogą być przekształcane na łańcuchy w produkcie IBM SPSS Modeler przy użyciu węzła rekodowania.

Zmienna przewidywana. W modelach klasyfikacji ODM tylko jedna zmienna może być wybrana jako wyjściowa (przewidywana).

Nazwa modelu. Począwszy od wersji Oracle 11gR1 nazwa `unique` jest słowem kluczowym i nie może być używana jako nazwa modelu niestandardowego.

Zmienna unikalna. Określa zmienną używaną w celu unikalnej identyfikacji każdej obserwacji. Na przykład może to być zmienna identyfikacyjna, taka jak *CustomerID*. Produkt IBM SPSS Modeler nakłada ograniczenie, że zmienna kluczowa musi być liczbowa.

Uwaga: ta zmienna jest opcjonalna dla wszystkich węzłów Oracle za wyjątkiem Oracle Adaptive Bayes, Oracle O-Cluster i Oracle Apriori.

Komentarze ogólne

- Produkt IBM SPSS Modeler nie udostępnia opcji importu/eksportu kodu PMML dla modeli utworzonych przez rozwiązanie Oracle Data Mining.
- Ocenianie modelu zawsze odbywa się w ODM. Jeśli dane pochodzą z produktu IBM SPSS Modeler lub muszą być w nim przygotowywane, wówczas konieczne może być wczytanie zestawu danych do tabeli tymczasowej.
- W produkcie IBM SPSS Modeler udostępniana jest zwykle tylko jedna predykcja i powiązane z nią prawdopodobieństwo lub współczynnik ufności.
- Produkt IBM SPSS Modeler ogranicza liczbę zmiennych przeznaczonych do użycia podczas budowania i oceniania modeli do 1000.
- Produkt IBM SPSS Modeler może oceniać modele ODM ze strumieni opublikowanych w celu wykonania przez produkt IBM SPSS Modeler Solution Publisher.

Modele Oracle — opcje serwera

Określ połączenie Oracle, które będzie używane w celu ładowania danych na potrzeby modelowania. W razie potrzeby można wybrać połączenie na karcie Serwer dla każdego węzła modelowania, aby przesłonić domyślne połączenie z Oracle określone w oknie dialogowym Aplikacje pomocnicze. Więcej informacji można znaleźć w temacie “Aktywacja integracji z produktem Oracle” na stronie 28.

Komentarze

- Połączenie używane na potrzeby modelowania może być takie samo, jak połączenie używane w węźle źródłowym dla strumienia. Na przykład może istnieć strumień, który uzyskuje dostęp do danych z jednej z baz danych Oracle, pobiera dane do produktu IBM SPSS Modeler w celu kasowania lub w celu przeprowadzenia innych manipulacji, a następnie przesyła te dane do innej bazy danych Oracle na potrzeby modelowania.
- Nazwa źródła danych ODBC jest osadzona w każdym strumieniu IBM SPSS Modeler. Jeśli strumień utworzony w jednym hoście zostanie wykonany na innym hoście, wówczas nazwa źródła danych musi być taka sama na każdym hoście. Alternatywnie inne źródło danych można wybrać na karcie Serwer w każdym węźle źródłowym lub węźle modelowania.

Koszty błędnej klasyfikacji

W niektórych kontekstach pewne błędy są bardziej kosztowne od innych. Przykładowo bardziej kosztowne może być sklasyfikowanie osób składających wnioski kredytowe z wysokim poziomem ryzyka jako osób z niskim poziomem ryzyka (jeden rodzaj błędu) niż sklasyfikowanie osób składających wnioski z niskim poziomem ryzyka jako osób z wysokim poziomem ryzyka (inny rodzaj błędu). Koszty błędnej klasyfikacji umożliwiają określenie względnej ważności różnych rodzajów błędów predykcji.

Kosztami błędnej klasyfikacji zwykle są wagi zastosowane do określonych danych wynikowych. Wagi te są uwzględniane w modelu i rzeczywiście mogą zmienić predykcję (jako sposób ochrony przed kosztownymi błędami).

Z wyjątkiem modeli C5.0 koszty błędnej klasyfikacji nie mają zastosowania podczas oceniania modelu i nie są brane pod uwagę podczas rangowania lub porównywania modeli za pomocą węzła Auto Klasyfikacja, wykresu ewaluacyjnego lub węzła analizy. Model, który uwzględnia koszty, może nie wygenerować mniejszej liczby błędów niż ten, który ich nie uwzględnia, i może nie mieć wyższej rangi pod względem całkowitej dokładności, ale prawdopodobnie lepiej sprawdzi się w warunkach praktycznych, ponieważ generuje błędy *mniej kosztowne*.

Macierz kosztów przedstawia koszty dla każdej możliwej kombinacji przewidywanej kategorii rzeczywistej. Domyślnie wszystkie koszty błędnej klasyfikacji są ustawione na wartość 1,0. Aby wprowadzić niestandardowe wartości kosztów, należy wybrać opcję **Stosuj koszty błędnej klasyfikacji** i wprowadzić do macierzy kosztów niestandardowe wartości.

Aby zmienić koszt błędnej klasyfikacji, należy zaznaczyć komórkę odpowiadającą odpowiedniej kombinacji wartości przewidywanych i rzeczywistych, usunąć istniejącą zawartość komórki i wprowadzić do niej żądany koszt. Koszty nie są automatycznie symetryczne. Przykładowo, jeśli koszt błędnej klasyfikacji *A* jako *B* zostanie ustawiony na 2,0, to koszt błędnej klasyfikacji *B* jako *A* nadal będzie miał domyślną wartość 1,0, chyba że zostanie ona również jawnie zmieniona.

Uwaga: tylko w przypadku modelu drzewa decyzyjnego możliwe jest określenie kosztów już na etapie budowania modelu.

Oracle Naive Bayes

Naive Bayes to dobrze znany algorytm stosowany w przypadku problemów z klasyfikacją. Model jest określany jako *naiwny*, ponieważ traktuje wszystkie proponowane zmienne predykcji jako niezależne od siebie. Naive Bayes to szybki, skalowalny algorytm, który oblicza prawdopodobieństwa warunkowe kombinacji atrybutów i atrybutu przewidywanego. Na podstawie danych uczących wyznaczane jest niezależne prawdopodobieństwo. To prawdopodobieństwo określa wiarygodność każdej klasy przewidywanej z uwzględnieniem wystąpienia każdej kategorii wartości z poszczególnych zmiennych wejściowych.

- Walidacja krzyżowa jest używana do testowania dokładności modelu względem tych samych danych, które były używane w celu budowania modelu. Jest to szczególnie użyteczne, gdy liczba obserwacji dostępnych do budowania modelu jest niewielka.
- Wynik tworzenia modelu można przeglądać w formacie macierzy. Liczby w macierzy są warunkowymi prawdopodobieństwami, które dotyczą przewidywanych klas (kolumny) oraz kombinacji zmienna predyktora-wartość (wiersze).

Opcje modelu Naive Bayes

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu, w przypadkach gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Użyj danych podzielonych na podzbiory. Jeśli zdefiniowano zmienną dzielącą na podzbiory, ta opcja umożliwia użycie podczas budowania modelu wyłącznie danych z podzbioru uczącego.

Zmienna unikalna. Określa zmienną używaną w celu unikalnej identyfikacji każdej obserwacji. Na przykład może to być zmienna identyfikacyjna, taka jak *CustomerID*. Produkt IBM SPSS Modeler nakłada ograniczenie, że zmienna kluczowa musi być liczbowa.

Uwaga: ta zmienna jest opcjonalna dla wszystkich węzłów Oracle za wyjątkiem Oracle Adaptive Bayes, Oracle O-Cluster i Oracle Apriori.

Automatyczne przygotowanie danych. (Dotyczy tylko 11g) Ta opcja włącza (stan domyślny) lub wyłącza tryb automatycznego przygotowywania danych produktu Oracle Data Mining. Jeśli to pole wyboru jest zaznaczone, wówczas ODM automatycznie przeprowadza transformacje danych wymagane przez algorytm. Więcej informacji zawiera dokumentacja *Koncepcje w programie Oracle Data Mining*.

Opcje zaawansowane Naive Bayes

Podczas budowania modelu pojedyncze wartości atrybutów predyktora i pary wartości są ignorowane, chyba że istnieje wystarczająca liczba wystąpień konkretnej wartości lub pary w danych uczących. Progi ignorowania wartości są określone jako frakcje na podstawie liczby rekordów w danych uczących. Modyfikacja tych progów może zmniejszyć szum i poprawić zdolność modelu do uogólniania innych zestawów danych.

- **Pojedyncza wartość graniczna.** Określa wartość graniczną dla konkretnej wartości atrybutu predyktora. Liczba wystąpień danej wartości musi być równa lub musi przekraczać określoną frakcję, ponieważ w przeciwnym wypadku wartość będzie ignorowana.
- **Wartość graniczna parami.** Określa wartość graniczną dla konkretnej pary atrybut i wartość predyktora. Liczba wystąpień danej pary wartości musi być równa lub musi przekraczać określoną frakcję, ponieważ w przeciwnym wypadku para będzie ignorowana.

Prawdopodobieństwo predykcji. Dzięki tej opcji model może uwzględniać prawdopodobieństwo właściwej predykcji dla możliwego rezultatu zmiennej przewidywanej. W celu aktywacji tej funkcji należy wybrać opcję **Wybierz**, kliknąć przycisk **Określ**, wybrać jeden z możliwych wyników i kliknąć opcję **Wstaw**.

Użyj zbioru predykcji. Generuje tabelę zawierającą wszystkie możliwe wyniki dla wszystkich możliwych rezultatów zmiennej przewidywanej.

Oracle Adaptive Bayes

Algorytm Adaptive Bayes Network (ABN) tworzy klasyfikatory sieci Bayesa, używając Minimum Description Length (MDL) i automatycznego wyboru predyktora. Algorytm ABN działa poprawnie w sytuacjach, gdy algorytm Naive Bayes działa słabo, i działa co najmniej tak samo dobrze w większości innych sytuacji, chociaż działanie może być wolniejsze. Algorytm ABN umożliwia zbudowanie trzech typów zaawansowanych modeli Bayesowskich, w tym uproszczonego drzewa decyzyjnego (jeden predyktor), przyciętego Naive Bayes i wzmocnionych modeli z wieloma predyktorami.

Uwaga: Algorytm Oracle Adaptive Bayes został wycofany w Oracle 12C i nie jest obsługiwany w produkcie IBM SPSS Modeler, gdy używana jest wersja Oracle 12C. Patrz http://docs.oracle.com/database/121/DMPRG/release_changes.htm#DMPRG726.

Wygenerowane modele

W trybie budowania z jednym predyktorem ABN generuje uproszczone drzewo decyzyjne oparte na zestawie reguł dostępnych do odczytu dla człowieka, które umożliwiają użytkownikowi biznesowemu i analitykowi zrozumienie podstaw predykcji modelu, a także działanie lub wyjaśnienie tych zasad innym. To może być znacząca zaleta w porównaniu z modelami Naive Bayes i modelami z wieloma predyktorami. Te reguły można przeglądać jak standardowy zestaw reguł w produkcie IBM SPSS Modeler. Prosty zestaw reguł może wyglądać jak poniżej:

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

Przycięte modele Naive Bayes i modele z wieloma składnikami nie mogą być przeglądane w IBM SPSS Modeler.

Opcje modelu Adaptive Bayes

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu, w przypadkach gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Użyj danych podzielonych na podzbiory. Jeśli zdefiniowano zmienną dzielącą na podzbiory, ta opcja umożliwia użycie podczas budowania modelu wyłącznie danych z podzbioru uczącego.

Zmienna unikalna. Określa zmienną używaną w celu unikalnej identyfikacji każdej obserwacji. Na przykład może to być zmienna identyfikacyjna, taka jak *CustomerID*. Produkt IBM SPSS Modeler nakłada ograniczenie, że zmienna kluczowa musi być liczbowa.

Uwaga: ta zmienna jest opcjonalna dla wszystkich węzłów Oracle za wyjątkiem Oracle Adaptive Bayes, Oracle O-Cluster i Oracle Apriori.

Typ modelu

Podczas budowania modelu do wyboru dostępne są trzy różne tryby.

- **Wiele predyktorów.** Buduje i porównuje szereg modeli, w tym model NB oraz modele prawdopodobieństwa produktu z jednym predyktorem lub wieloma predyktorami. Ten tryb jest najbardziej kompleksowy i zwykle w związku z tym obliczenie trwa najdłużej. Reguły są generowane tylko wówczas, gdy okaże się, że model z jednym predyktorem jest najlepszy. Jeśli zostanie wybrany model z wieloma predyktorami lub model NB, wówczas żadne reguły nie zostaną wygenerowane.
- **Jeden predyktor.** Tworzy uproszczone drzewo decyzyjne na podstawie zestawu reguł. Każda reguła zawiera warunek, a także prawdopodobieństwa powiązane z każdym wynikiem. Reguły wzajemnie się wykluczają i są udostępniane w formacie dostępnym do odczytu dla człowieka, co może być znaczącą zaletą w porównaniu z modelami Naive Bayes i modelami z wieloma predyktorami.
- **Naive Bayes.** Buduje pojedynczy model NB i porównuje go z próbą globalną a priori (rozkład wartości przewidywanych w próbie globalnej). Model NB jest generowany jako wynik tylko wtedy, gdy okaże się, że jest lepszym predyktorem wartości przewidywanych niż globalne prawdopodobieństwo a priori. W przeciwnym wypadku żaden model nie zostanie wygenerowany jako wynik.

Opcje zaawansowane Adaptive Bayes

Ogranicz czas wykonywania. Wybierz tę opcję, aby określić maksymalny czas budowania w minutach. Dzięki temu możliwe jest uzyskiwanie modeli w krótszym czasie, mimo że modele wynikowe mogą być mniej dokładne. Po osiągnięciu każdego kamienia milowego w procesie modelowania algorytm sprawdza (przed wykonaniem dalszych operacji), czy będzie mógł osiągnąć następny kamień milowy w określonym czasie, a następnie zwraca najlepszy model dostępny po osiągnięciu limitu.

Maksimum predyktorów. Ta opcja umożliwia ograniczenie złożoności modelu i poprawę wydajności poprzez ograniczenie liczby używanych predyktorów. Predyktorom przyznawane są rangi na podstawie miary MDL ich korelacji ze zmienną przewidywaną jako miarą wiarygodności tego, że zostaną uwzględnione w modelu.

Maksimum predyktorów Naive Bayes. Ta opcja określa maksymalną liczbę predyktorów, które będą używane w modelu Naive Bayes.

Oracle Support Vector Machine (SVM)

Algorytm Support Vector Machine (SVM) to algorytm klasyfikacji i regresji, który wykorzystuje teorię uczenia maszynowego w celu maksymalizacji predykcyjnej dokładności bez przeuczania danych. SVM wykorzystuje opcjonalną transformację nieliniową danych uczących, po czym następuje wyszukiwanie równań regresji w danych po transformacji w celu rozdzielenia klas (dla jakościowych zmiennych przewidywanych) lub dopasowania zmiennej przewidywanej (dla ciągłych zmiennych przewidywanych). Implementacja SVM Oracle umożliwia budowanie modeli przy użyciu jednego z dwóch dostępnych algorytmów — liniowego i gaussowskiego. Algorytm liniowy pomija w całości transformację nieliniową, dlatego model wynikowy jest modelem regresji.

Więcej informacji zawiera dokumentacja *Oracle Data Mining Application Developer's Guide* oraz *Oracle Data Mining Concepts*.

Opcje modelu Oracle SVM

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu, w przypadkach gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Zmienna unikalna. Określa zmienną używaną w celu unikalnej identyfikacji każdej obserwacji. Na przykład może to być zmienna identyfikacyjna, taka jak *CustomerID*. Produkt IBM SPSS Modeler nakłada ograniczenie, że zmienna kluczowa musi być liczbowa.

Uwaga: ta zmienna jest opcjonalna dla wszystkich węzłów Oracle za wyjątkiem Oracle Adaptive Bayes, Oracle O-Cluster i Oracle Apriori.

Automatyczne przygotowanie danych. (Dotyczy tylko 11g) Ta opcja włącza (stan domyślny) lub wyłącza tryb automatycznego przygotowywania danych produktu Oracle Data Mining. Jeśli to pole wyboru jest zaznaczone, wówczas ODM automatycznie przeprowadza transformacje danych wymagane przez algorytm. Więcej informacji zawiera dokumentacja *Koncepcje w programie Oracle Data Mining*.

Aktywna nauka. Ta opcja zapewnia sposób radzenia sobie z dużymi zestawami budowania. W przypadku aktywnej nauki algorytm tworzy początkowy model na podstawie niewielkiej próby, następnie stosuje go do kompletnego zestawu danych uczących, a dopiero później stopniowo aktualizuje próbę i model na podstawie wyników. Taki cykl jest powtarzany, aż model osiągnie zbieżność w danych uczących lub zostanie osiągnięta maksymalna liczba wektorów obsługi.

Funkcja algorytmu domyślnego. Wybierz opcję **Liniowy** lub **Gausa** albo pozostaw domyślną wartość **Określona przez system**, aby umożliwić systemowi wybór najlepiej dopasowanego algorytmu. Algorytmy gaussowskie można nauczyć bardziej złożonych zależności, ale zwykle obliczenia trwają dłużej. Na początku można użyć algorytmu liniowego, a następnie wypróbować algorytm gaussowski, jeśli liniowy nie znajdzie dobrego dopasowania. Dużo większe jest prawdopodobieństwo takiej sytuacji w przypadku modelu regresji, gdzie wybór algorytmu ma większe znaczenie. Ponadto należy zwrócić uwagę na to, że modele SVM zbudowane przy użyciu algorytmu gaussowskiego nie mogą być przeglądane w produkcie IBM SPSS Modeler. Modele zbudowane za pomocą algorytmu liniowego mogą być przeglądane w produkcie IBM SPSS Modeler w taki sam sposób, jak standardowe modele regresji.

Metoda normalizacji. Określa metodę normalizacji dla ciągłych zmiennych wejściowych i zmiennych przewidywanych. Do wyboru dostępne są opcje **Statystyki z**, **Min.-Maks.** oraz **Brak**. Oracle przeprowadza normalizację automatycznie, jeśli zaznaczone jest pole wyboru **Automatyczne przygotowanie danych**. Usunięcie zaznaczenia tego pola wyboru powoduje, że metodę normalizacji należy wybierać ręcznie.

Opcje zaawansowane Oracle SVM

Rozmiar pamięci podręcznej. Określa w bajtach rozmiar pamięci podręcznej, która będzie używana do zapisywania obliczanych algorytmów podczas operacji budowania. Zgodnie z oczekiwaniem większe pamięci podręczne zapewniają krótsze czasy budowania. Domyślną wartością jest 50 MB.

Tolerancja zbieżności. Określa wartość tolerancji, która jest dozwolona przed zakończeniem na potrzeby budowania modelu. Wartość musi mieścić się w przedziale od 0 do 1. Wartością domyślną jest 0,001. Większe wartości wykazują tendencję do szybszego budowania, ale mniej dokładnych modeli.

Określ odchylenie standardowe. Określa parametr odchylenia standardowego używany przez algorytm gaussowski. Ten parametr określa kompromis między złożonością modelu a zdolnością do uogólniania do innych zestawów danych (przeuczenie i niedouczenie). Wyższe wartości odchylenia standardowego sprzyjają niedouczeniu. Domyślnie ten parametr jest szacowany z danych uczących.

Określ Epsilon. Dotyczy tylko modeli regresji i określa wartość przedziału dozwolonego błędu podczas budowania modeli, w których często stosowane są wartości epsilon. Innymi słowy, odróżnia niewielkie błędy (które są ignorowane) od dużych błędów (które nie są). Wartość musi mieścić się w przedziale od 0 do 1. Domyślnie jest szacowana z danych uczących.

Określ czynnik złożoności. Określa czynnik złożoności, czyli kompromis między błędem modelu (mierzonym względem danych uczących) a złożonością modelu w celu uniknięcia przeuczenia i niedouczenia. Wyższe wartości stosują większą karę w przypadku błędu, przy czym istnieje zwiększone ryzyko przeuczenia; niższe wartości stosują mniejszą karę w razie błędu i mogą prowadzić do niedouczenia.

Określ wskaźnik wartości odstających. Określa żądany współczynnik wartości odstających w danych uczących. Obowiązuje tylko w przypadku modeli One-Class SVM. Nie może być używana z ustawieniem **Określ czynnik złożoności**.

Prawdopodobieństwo predykcji. Dzięki tej opcji model może uwzględniać prawdopodobieństwo właściwej predykcji dla możliwego rezultatu zmiennej przewidywanej. W celu aktywacji tej funkcji należy wybrać opcję **Wybierz**, kliknąć przycisk **Określ**, wybrać jeden z możliwych wyników i kliknąć opcję **Wstaw**.

Użyj zbioru predykcji. Generuje tabelę zawierającą wszystkie możliwe wyniki dla wszystkich możliwych rezultatów zmiennej przewidywanej.

Opcje wag dla Oracle SVM

W modelu klasyfikacji stosowanie wag pozwala określać względną ważność różnych możliwych wartości przewidywanych. Taki sposób postępowania może być użyteczny, na przykład jeśli punkty danych w danych uczących nie są rozłożone realistycznie między kategorie. Wagi umożliwiają obciążenie modelu, dzięki czemu możliwe jest skompensowanie tych kategorii, które są gorzej reprezentowane w danych. Zwiększenie wagi wartości przewidywanej powinno spowodować wzrost procentu poprawnych predykcji dla tej kategorii.

Istnieją trzy metody ustawienia wag:

- **W oparciu o dane uczące.** Jest to ustawienie domyślne. Wagi określane są w oparciu o względne częstości kategorii w danych uczących.
- **Równe dla wszystkich klas.** Wagi dla wszystkich kategorii są definiowane jako $1/k$, gdzie k to liczba kategorii zmiennej przewidywanej.
- **Użytkownika.** Użytkownik może określić własne wagi. Wartości początkowe dla wag są ustawiane jako równe dla wszystkich klas. Można dostosować wagi dla poszczególnych kategorii, ustawiając wartości zdefiniowane przez użytkownika. Aby ustawić określoną wagę dla kategorii, należy wybrać komórkę Waga w tabeli odpowiadającej żądanej kategorii, usunąć zawartość komórki i wprowadzić żadaną wartość.

Suma wag dla wszystkich kategorii powinna wynosić 1,0. Jeśli suma nie wynosi 1,0, wyświetlane jest ostrzeżenie z opcją automatycznego znormalizowania wartości. To automatyczne dostosowanie zachowuje proporcje we wszystkich kategoriach, wymuszając ograniczenie wagi. Takie dostosowanie można przeprowadzić w dowolnym czasie, klikając przycisk **Normalizuj**. Aby w tabeli ponownie ustawić jednakowe wartości dla wszystkich kategorii, należy kliknąć przycisk **Wyrównaj**.

Uogólnione modele liniowe Oracle (GLM)

(Dotyczy tylko 11g) Uogólnione modele liniowe rozluźniają restrykcyjne założenia pochodzące z modeli liniowych. Są to między innymi założenia, że zmienna przewidywana ma rozkład normalny, a wpływ predyktorów na zmienną przewidywaną jest liniowy z natury. Uogólniony model liniowy jest odpowiedni w przypadku predykcji, w których rozkład zmiennej przewidywanej może być inny niż normalny, np. wielomianowy lub Poissona. Uogólniony model liniowy jest użyteczny w przypadkach, w których zależność lub połączenie między predyktorami i zmienną przewidywaną mogą być nieliniowe.

Więcej informacji zawiera dokumentacja *Oracle Data Mining Application Developer's Guide* oraz *Oracle Data Mining Concepts*.

Opcje modelu Oracle GLM

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu, w przypadkach gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Zmienna unikalna. Określa zmienną używaną w celu unikalnej identyfikacji każdej obserwacji. Na przykład może to być zmienna identyfikacyjna, taka jak *CustomerID*. Produkt IBM SPSS Modeler nakłada ograniczenie, że zmienna kluczowa musi być liczbowa.

Uwaga: ta zmienna jest opcjonalna dla wszystkich węzłów Oracle za wyjątkiem Oracle Adaptive Bayes, Oracle O-Cluster i Oracle Apriori.

Automatyczne przygotowanie danych. (Dotyczy tylko 11g) Ta opcja włącza (stan domyślny) lub wyłącza tryb automatycznego przygotowywania danych produktu Oracle Data Mining. Jeśli to pole wyboru jest zaznaczone, wówczas ODM automatycznie przeprowadza transformacje danych wymagane przez algorytm. Więcej informacji zawiera dokumentacja *Koncepcje w programie Oracle Data Mining*.

Metoda normalizacji. Określa metodę normalizacji dla ciągłych zmiennych wejściowych i zmiennych przewidywanych. Do wyboru dostępne są opcje **Statystyki z**, **Min.-Maks.** oraz **Brak**. Oracle przeprowadza normalizację automatycznie, jeśli zaznaczone jest pole wyboru **Automatyczne przygotowanie danych**. Usunięcie zaznaczenia tego pola wyboru powoduje, że metodę normalizacji należy wybierać ręcznie.

Traktowanie braków danych. Określa sposób przetwarzania brakujących wartości w danych wejściowych:

- **Zastąp średnią lub dominantą** — zastępuje wartości brakujące atrybutów liczbowych wartością średnią i zastępuje wartości brakujące atrybutów jakościowych dominantą.
- **Używaj tylko kompletnych rekordów** — ignoruje rekordy z brakującymi wartościami.

Opcje zaawansowane Oracle GLM

Użyj wag wierszy. Zaznacz to pole wyboru, aby aktywować sąsiednią listę rozwijaną, z której można wybrać kolumnę zawierającą czynnik wagi dla wierszy.

Zapisz surowe dane diagnostyczne w tabeli. Zaznacz to pole wyboru, aby aktywować sąsiednie pole tekstowe, w którym można wprowadzić nazwę tabeli, która będzie zawierać dane diagnostyczne na poziomie wiersza.

Poziom istotności współczynnika. Stopień pewności od 0,0 do 1,0 oznacza, że wartość przewidywana dla zmiennej przewidywanej będzie należeć do przedziału ufności wyliczonego przez model. Granice przedziału ufności są zwracane ze statystykami dot. współczynnika.

Kategoria odniesienia dla zmiennej przewidywanej. Wybierz opcję **Użytkownika**, aby wybrać wartość dla zmiennej przewidywanej, która będzie używana jako kategoria odniesienia, albo pozostaw wartość domyślną **Automatycznie**.

Regresja grzbietowa. Regresja grzbietowa to technika, które kompensuje sytuację, gdy stopień korelacji w zmiennych jest zbyt wysoki. Można użyć opcji **Automatycznie**, aby zezwolić algorytmowi na kontrolowanie użycia tej techniki, albo można kontrolować ją ręcznie przy użyciu opcji **Wyłącz** i **Włącz**. Jeśli zdecydujesz się włączyć regresję grzbietową ręcznie, możesz zastąpić wartość domyślną systemu dla parametru grzbietu, wprowadzając wartość do sąsiedniego pola.

Utwórz VIF dla regresji grzbietowej. Zaznacz to pole wyboru, jeśli zamierzasz wygenerować statystyki Variance Inflation Factor (VIF), gdy regresja grzbietowa jest używana dla regresji liniowej.

Prawdopodobieństwo predykcji. Dzięki tej opcji model może uwzględniać prawdopodobieństwo właściwej predykcji dla możliwego rezultatu zmiennej przewidywanej. W celu aktywacji tej funkcji należy wybrać opcję **Wybierz**, kliknąć przycisk **Określ**, wybrać jeden z możliwych wyników i kliknąć opcję **Wstaw**.

Użyj zbioru predykcji. Generuje tabelę zawierającą wszystkie możliwe wyniki dla wszystkich możliwych rezultatów zmiennej przewidywanej.

Opcje wag dla Oracle GLM

W modelu klasyfikacji stosowanie wag pozwala określać względną ważność różnych możliwych wartości przewidywanych. Taki sposób postępowania może być użyteczny, na przykład jeśli punkty danych w danych uczących nie są rozłożone realistycznie między kategorie. Wagi umożliwiają obciążenie modelu, dzięki czemu możliwe jest skompensowanie tych kategorii, które są gorzej reprezentowane w danych. Zwiększenie wagi wartości przewidywanej powinno spowodować wzrost procentu poprawnych predykcji dla tej kategorii.

Istnieją trzy metody ustawienia wag:

- **W oparciu o dane uczące.** Jest to ustawienie domyślne. Wagi określane są w oparciu o względne częstości kategorii w danych uczących.
- **Równe dla wszystkich klas.** Wagi dla wszystkich kategorii są definiowane jako $1/k$, gdzie k to liczba kategorii zmiennej przewidywanej.
- **Użytkownika.** Użytkownik może określić własne wagi. Wartości początkowe dla wag są ustawiane jako równe dla wszystkich klas. Można dostosować wagi dla poszczególnych kategorii, ustawiając wartości zdefiniowane przez użytkownika. Aby ustawić określoną wagę dla kategorii, należy wybrać komórkę Waga w tabeli odpowiadającej żądanej kategorii, usunąć zawartość komórki i wprowadzić żądaną wartość.

Suma wag dla wszystkich kategorii powinna wynosić 1,0. Jeśli suma nie wynosi 1,0, wyświetlane jest ostrzeżenie z opcją automatycznego znormalizowania wartości. To automatyczne dostosowanie zachowuje proporcje we wszystkich kategoriach, wymuszając ograniczenie wagi. Takie dostosowanie można przeprowadzić w dowolnym czasie, klikając przycisk **Normalizuj**. Aby w tabeli ponownie ustawić jednakowe wartości dla wszystkich kategorii, należy kliknąć przycisk **Wyrównaj**.

Oracle Decision Tree

Oracle Data Mining oferuje predyktor Drzewo decyzyjne oparty na popularnym algorytmie klasyfikacji i drzewa regresji. Model drzewa decyzyjnego ODM zawiera kompletne informacje dotyczące poszczególnych węzłów, w tym ufność, informacje o obsłudze i kryterium podziału. Możliwe jest wyświetlenie pełnej reguły dla każdego węzła, a ponadto dla każdego węzła dostępny jest substytut atrybutu, który można wykorzystać jako zastępczy w przypadku zastosowania modelu względem obserwacji, w której brakuje wartości.

Drzewa decyzyjne są popularne, ponieważ znajdują wiele zastosowań, można je łatwo stosować i zrozumieć. Drzewa decyzyjne przesiewają każdy potencjalny atrybut wejściowy w poszukiwaniu najlepszego „rozdzielacza”, czyli punktu podziału (na przykład $AGE > 55$), który spowoduje podział dalszych rekordów danych na bardziej jednorodną populację. Po każdej decyzji o podziale ODM powtarza proces, powiększając całe drzewo i tworząc liście końcowe, które reprezentują podobne populacje rekordów, elementów lub ludzi. Gdy spoglądamy z węzła głównego drzewa (na przykład populacji łącznej), drzewa decyzyjne udostępniają reguły czytelne dla ludzi, które zawierają instrukcje IF A, then B. Te reguły drzew decyzyjnych zapewniają również obsługę i ufność dla każdego węzła drzewa.

Sieci Adaptive Bayes Networks również udostępniają krótkie proste reguły, które mogą być użyteczne w wyjaśnianiu poszczególnych predykcji, ale drzewa decyzyjne zapewniają pełne reguły Oracle Data Mining dla każdej decyzji o podziale. Drzewa decyzyjne są również użyteczne w przypadku opracowywania szczegółowych profili najlepszych klientów, zdrowych pacjentów, czynników powiązanych z oszustwem itp.

Opcje modelu drzewa decyzyjnego

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu, w przypadkach gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Zmienna unikalna. Określa zmienną używaną w celu unikalnej identyfikacji każdej obserwacji. Na przykład może to być zmienna identyfikacyjna, taka jak *CustomerID*. Produkt IBM SPSS Modeler nakłada ograniczenie, że zmienna kluczowa musi być liczbowa.

Uwaga: ta zmienna jest opcjonalna dla wszystkich węzłów Oracle za wyjątkiem Oracle Adaptive Bayes, Oracle O-Cluster i Oracle Apriori.

Automatyczne przygotowanie danych. (Dotyczy tylko 11g) Ta opcja włącza (stan domyślny) lub wyłącza tryb automatycznego przygotowywania danych produktu Oracle Data Mining. Jeśli to pole wyboru jest zaznaczone, wówczas ODM automatycznie przeprowadza transformacje danych wymagane przez algorytm. Więcej informacji zawiera dokumentacja *Koncepcje w programie Oracle Data Mining*.

Metryka zanieczyszczenia. Określa metrykę, która jest używana do znalezienia najlepszego pytania testowego w przypadku podziału danych na każdym węźle. Najlepszym rozdzielaczem i najlepszą wartością podziału są te wartości,

które powodują największy przyrost jednorodności wartości przewidywanej dla jednostek w węźle. Jednorodność jest mierzona zgodnie z metryką. Obsługiwane są metryki **gini** i **entropia**.

Opcje zaawansowane drzewa decyzyjnego

Maksymalna głębokość. Ustawia maksymalną głębokość modelu drzewa przeznaczonego do zbudowania.

Minimalna wielkość procentowa rekordów w węźle. Ustawia procent minimalnej liczby rekordów na węzeł.

Minimalna wielkość procentowa rekordów w podzbiorze. Ustawia minimalną liczbę rekordów w węźle nadrzędnym, co jest wyrażone jako procent łącznej liczby rekordów używanych do uczenia modelu. Jeśli liczba rekordów jest poniżej tej wartości procentowej, próba podziału nie jest podejmowana.

Minimalna liczba rekordów w węźle. Ustawia minimalną liczbę zwracanych rekordów.

Minimalna liczba rekordów w podzbiorze. Ustawia minimalną liczbę rekordów w węźle nadrzędnym, co jest wyrażone jako wartość. Jeśli liczba rekordów jest poniżej tej wartości, próba podziału nie jest podejmowana.

Identyfikator reguły. Jeśli ta opcja jest wybrana, wówczas uwzględnia w modelu łańcuch w celu identyfikacji węzła w drzewie, w którym wykonywany jest konkretny podział.

Prawdopodobieństwo predykcji. Dzięki tej opcji model może uwzględniać prawdopodobieństwo właściwej predykcji dla możliwego rezultatu zmiennej przewidywanej. W celu aktywacji tej funkcji należy wybrać opcję **Wybierz**, kliknąć przycisk **Określ**, wybrać jeden z możliwych wyników i kliknąć opcję **Wstaw**.

Użyj zbioru predykcji. Generuje tabelę zawierającą wszystkie możliwe wyniki dla wszystkich możliwych rezultatów zmiennej przewidywanej.

Oracle O-Cluster

Algorytm Oracle O-Cluster identyfikuje naturalnie występujące grupowania w populacji danych. Grupowanie przez dzielenie ortogonalne (O-Cluster) to zastrzeżony algorytm grupowania Oracle, który tworzy hierarchiczny model grupowania oparty na siatce, czyli taki, który tworzy podziały równoległe do osi (ortogonalne) w przestrzeni atrybutów wejściowych. Algorytm działa rekurencyjnie. Wynikowa struktura hierarchiczna reprezentuje nieregularną siatkę, która mozaikowo dzieli przestrzeń atrybutów na skupienia.

Algorytm O-Cluster obsługuje atrybuty liczbowe i jakościowe, a ODM automatycznie wybiera najlepsze definicje skupień. Produkt ODM udostępni szczegółowe informacje o skupieniach, reguły skupień, wartości środków skupień i może być używany w celu oceny populacji na podstawie ich przynależności do skupień.

Opcje modelu O-Cluster

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu, w przypadkach gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Zmienna unikalna. Określa zmienną używaną w celu unikalnej identyfikacji każdej obserwacji. Na przykład może to być zmienna identyfikacyjna, taka jak *CustomerID*. Produkt IBM SPSS Modeler nakłada ograniczenie, że zmienna kluczowa musi być liczbową.

Uwaga: ta zmienna jest opcjonalna dla wszystkich węzłów Oracle za wyjątkiem Oracle Adaptive Bayes, Oracle O-Cluster i Oracle Apriori.

Automatyczne przygotowanie danych. (Dotyczy tylko 11g) Ta opcja włącza (stan domyślny) lub wyłącza tryb automatycznego przygotowywania danych produktu Oracle Data Mining. Jeśli to pole wyboru jest zaznaczone, wówczas ODM automatycznie przeprowadza transformacje danych wymagane przez algorytm. Więcej informacji zawiera dokumentacja *Koncepcje w programie Oracle Data Mining*.

Maksymalna liczba grup. Ustawia maksymalną liczbę generowanych skupień.

Opcje zaawansowane O-Cluster

Maksymalny rozmiar buforu. Ustawia maksymalny rozmiar buforu.

Czułość. Ustawia ułamek, który określa szczytową gęstość wymaganą do rozdzielenia nowego skupienia. Ten ułamek jest powiązany z globalną gęstością jednostajną.

Oracle K-średnie

Algorytm Oracle K-średnie identyfikuje naturalnie występujące grupowania w populacji danych. Algorytm k-średnie jest algorytmem grupowania zależnym od odległości, który dzieli dane na wstępnie określoną liczbę skupień (przy założeniu, że istnieje wystarczająca liczba odrębnych obserwacji). Algorytmy oparte na odległości są zależne od metryki odległości (funkcji), dzięki której mierzą podobieństwo między punktami danych. Punkty danych są przypisane do najbliższego skupienia zgodnie z używaną metryką odległości. ODM udostępnia rozszerzoną wersję algorytmu K-średnie.

Algorytm K-średnie obsługuje skupienia hierarchiczne, obsługuje atrybuty liczbowe i jakościowe, a także dzieli populację na zdefiniowaną przez użytkownika liczbę skupień. Produkt ODM udostępnia szczegółowe informacje o skupieniach, reguły skupień, wartości środków skupień i może być używany w celu oceny populacji na podstawie ich przynależności do skupień.

Opcje modelu K-średnie

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu, w przypadkach gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Zmienna unikalna. Określa zmienną używaną w celu unikalnej identyfikacji każdej obserwacji. Na przykład może to być zmienna identyfikacyjna, taka jak *CustomerID*. Produkt IBM SPSS Modeler nakłada ograniczenie, że zmienna kluczowa musi być liczbową.

Uwaga: ta zmienna jest opcjonalna dla wszystkich węzłów Oracle za wyjątkiem Oracle Adaptive Bayes, Oracle O-Cluster i Oracle Apriori.

Automatyczne przygotowanie danych. (Dotyczy tylko 11g) Ta opcja włącza (stan domyślny) lub wyłącza tryb automatycznego przygotowywania danych produktu Oracle Data Mining. Jeśli to pole wyboru jest zaznaczone, wówczas ODM automatycznie przeprowadza transformacje danych wymagane przez algorytm. Więcej informacji zawiera dokumentacja *Koncepcje w programie Oracle Data Mining*.

Liczba grup. Ustawia liczbę wygenerowanych skupień

Funkcja odległości. Określa funkcję odległości, która będzie używana na potrzeby grupowania metodą k-średnich.

Kryterium podziału. Określa kryterium podziału, które będzie używane na potrzeby grupowania metodą k-średnich.

Metoda normalizacji. Określa metodę normalizacji dla ciągłych zmiennych wejściowych i zmiennych przewidywanych. Do wyboru dostępne są opcje **Statystyki z**, **Min.-Maks.** oraz **Brak**.

Opcje zaawansowane K-średnich

Iteracje. Ustawia liczbę iteracji algorytmu K-średnich.

Tolerancja zbieżności. Ustawia tolerancję zbieżności dla algorytmu K-średnich.

Liczba przedziałów. Określa liczbę przedziałów w histogramie atrybutów wygenerowanym przez algorytm K-średnich. Granice przedziałów dla każdego atrybutu są obliczane globalnie względem całego uczącego zbioru danych. Metoda kategoryzacji to metoda przedziałów równej szerokości. Wszystkie atrybuty mają tę samą liczbę przedziałów z wyjątkiem atrybutów o pojedynczej wartości, które mają tylko jeden przedział.

Zablokuj wzrost. Ustawia współczynnik wzrostu dla przydzielonej pamięci w taki sposób, aby zatrzymał dane skupień.

Minimalny procent obsługi atrybutu. Ustawia procent atrybutów, które muszą być inne niż null, ponieważ tylko wówczas atrybut zostanie uwzględniony w opisie reguły dla skupienia. Ustawienie wartości parametru na zbyt wysoką wartość w danych, w których brakuje wartości, może spowodować bardzo krótkie lub nawet puste reguły.

Oracle Nonnegative Matrix Factorization (NMF)

Algorytm Nonnegative Matrix Factorization (NMF) jest używany w celu redukcji dużego zestawu danych do reprezentatywnych atrybutów. Ten algorytm przypomina algorytm Principal Components Analysis (PCA) pod względem koncepcji, ale może obsługiwać znacznie większe ilości atrybutów i w modelu reprezentacji addytywnej. NMF to wydajny i nowoczesny algorytm eksploracji danych, który może być używany do różnych celów.

Z algorytmu NMF można korzystać w celu redukcji dużych ilości danych, np. danych tekstowych, do postaci mniejszych, bardziej rozproszonych reprezentacji, które zmniejszają wymiarowość danych (te same informacje mogą być zachowywane przy użyciu dużo mniejszej liczby zmiennych). Wyniki modeli NMF mogą być analizowane z użyciem technik uczenia nadzorowanego, np. SVM, a także technik uczenia nienadzorowanego, takich jak techniki grupowania. Oracle Data Mining wykorzystuje algorytmy NMF i SVM w celu eksploracji nieustrukturyzowanych danych tekstowych.

Opcje modelu NMF

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu, w przypadkach gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Zmienna unikalna. Określa zmienną używaną w celu unikalnej identyfikacji każdej obserwacji. Na przykład może to być zmienna identyfikacyjna, taka jak *CustomerID*. Produkt IBM SPSS Modeler nakłada ograniczenie, że zmienna kluczowa musi być liczbowa.

Uwaga: ta zmienna jest opcjonalna dla wszystkich węzłów Oracle za wyjątkiem Oracle Adaptive Bayes, Oracle O-Cluster i Oracle Apriori.

Automatyczne przygotowanie danych. (Dotyczy tylko 11g) Ta opcja włącza (stan domyślny) lub wyłącza tryb automatycznego przygotowywania danych produktu Oracle Data Mining. Jeśli to pole wyboru jest zaznaczone, wówczas ODM automatycznie przeprowadza transformacje danych wymagane przez algorytm. Więcej informacji zawiera dokumentacja *Koncepcje w programie Oracle Data Mining*.

Metoda normalizacji. Określa metodę normalizacji dla ciągłych zmiennych wejściowych i zmiennych przewidywanych. Do wyboru dostępne są opcje **Statystyki z**, **Min.-Maks.** oraz **Brak**. Oracle przeprowadza normalizację automatycznie, jeśli zaznaczone jest pole wyboru **Automatyczne przygotowanie danych**. Usunięcie zaznaczenia tego pola wyboru powoduje, że metodę normalizacji należy wybierać ręcznie.

Opcje zaawansowane NMF

Określona liczba predyktorów. Określa liczbę predyktorów do wyodrębnienia.

Wartość początkowa. Ustawia wartość początkową generatora liczb losowych dla algorytmu NMF.

Liczba iteracji. Ustawia liczbę iteracji algorytmu NMF.

Tolerancja zbieżności. Ustawia tolerancję zbieżności dla algorytmu NMF.

Pokaż wszystkie predyktory. Ta opcja powoduje wyświetlanie identyfikatora i ufności predyktora dla wszystkich predyktorów, a nie tylko wartości dla najlepszego predyktora.

Oracle Apriori

Algorytm Apriori wykrywa reguły asocjacyjne w danych. Na przykład, „jeśli klient kupił golarzkę i płyn po goleniu, to z ufnością 80% klient kupi krem do golenia”. Problem eksploracji asocjacji można rozłożyć na dwa problemy podrzędne:

- Znajdź wszystkie kombinacje elementów, czasami nazywane zbiorami elementów, dla których zakres obsługi (support) jest większy niż minimalny.
- Użyj częstych zbiorów elementów, aby wygenerować żądane reguły. Idea polega na tym, że jeśli na przykład ABC i BC występują często, wówczas reguła „A sugeruje BC” będzie prawdziwa, jeśli stosunek $\text{support}(ABC)$ do $\text{support}(BC)$ jest co najmniej tak duży jak minimalna ufność. Należy zwrócić uwagę na to, że reguła będzie miała minimalne wsparcie (support), ponieważ ABCD jest częste. Asocjacja ODM obsługuje tylko reguły z pojedynczym następnikiem (ABC sugeruje D).

Liczba częstych zbiorów elementów jest określona przez parametry minimalnej obsługi (support). Liczba generowanych reguł jest określana przez liczbę częstych zbiorów elementów i przez parametr ufności. Jeśli parametr ufności jest ustawiony na wartość zbyt wysoką, wówczas w modelu asocjacyjnym mogą pojawić się częste zbiory elementów, ale bez reguł.

ODM używa implementacji algorytmu Apriori opartej na kodzie SQL. Etapy generowania kandydatów i zliczania opcji obsługi (support) są realizowane przy użyciu zapytań SQL. Specjalne struktury danych w pamięci nie są używane. Zapytania SQL są specjalnie modyfikowane w celu wydajnego działania na serwerze bazy danych, przy użyciu różnych wskazówek.

Opcje zmiennych Apriori

Wszystkie węzły modelowania zawierają kartę Zmienne, na której można określić zmienne, które będą używane podczas budowania modelu.

Aby możliwe było zbudowanie modelu Apriori, konieczne jest określenie, które zmienne mają być używane jako elementy zainteresowania w modelowaniu asocjacyjnym.

Użyj ustawień węzła Typ. Ta opcja stanowi dla węzła instrukcję o konieczności użycia informacji o zmiennej z wcześniejszego węzła Typ. Jest to ustawienie domyślne.

Użyj ustawień niestandardowych. Ta opcja stanowi dla węzła instrukcję o konieczności użycia informacji o zmiennej określonych w tym miejscu, a nie w żadnym wcześniejszym węźle Typ. Po wybraniu tej opcji wypełnij pozostałe pola w oknie dialogowym, co jest zależne od tego, czy używany jest format transakcyjny.

Jeśli *nie* jest używany format transakcyjny, określ:

- **Zmienne wejściowe.** Wybierz jedną lub więcej zmiennych wejściowych. Działanie jest podobne jak w przypadku ustawienia roli zmiennej na wartość *Dane wejściowe* w węźle Typ.
- **Podział.** To pole umożliwi określenie zmiennej używanej do podziału danych na osobne próby do uczenia, testowania i walidacji podczas budowania modelu.

Jeśli format transakcyjny *jest* używany, określ:

Użyj formatu transakcyjnego. Użyj tej opcji, jeśli zamierzasz przekształcić dane wiersz na element na dane wiersz na obserwację.

Wybranie tej opcji powoduje zmianę elementów sterowania zmiennymi w dolnej części tego okna dialogowego:

Na potrzeby formatu transakcyjnego określ:

- **Identyfikator.** Należy wybrać zmienną identyfikacyjną z listy. Jako zmienna identyfikacyjna mogą być używane zmienne numeryczne lub symboliczne. Każda unikalna wartość tej zmiennej powinna wskazywać na określoną jednostkę analizy. Na przykład w aplikacji do obsługi koszyka zakupów każdy identyfikator może reprezentować jednego klienta. W przypadku aplikacji do analizy dzienników sieciowych każdy identyfikator może reprezentować komputer (wg adresu IP) lub użytkownika (wg danych logowania).
- **Zawartość.** Określ zmienne zawartości dla modelu. Ta zmienna zawiera element zainteresowania w modelowaniu asocjacyjnym.
- **Podział.** To pole umożliwia określenie zmiennej używanej do podziału danych na osobne próby do uczenia, testowania i walidacji podczas budowania modelu. Korzystając z jednej próby do utworzenia modelu oraz innej do testowania go, można uzyskać wiarygodne informacje o tym, jak dobrze model pozwala uogólnić większe zbiory danych, podobne do danych bieżących. Jeśli korzystając z węzłów Typ lub Partycja, zdefiniowano wiele zmiennych dzielących na podzbiory, na karcie Zmienne każdego węzła modelowania korzystającego z tego podziału na podzbiory należy wybrać jedną zmienną dzielącą na podzbiory. (Jeśli obecna jest tylko jedna zmienna dzieląca na podzbiory, jest ona automatycznie używana po aktywowaniu dzielenia). Należy również pamiętać, że aby zastosować wybrany podział w analizie, dzielenie musi być również włączone na karcie Opcje modelu danego węzła. (Usunięcie zaznaczenia tej opcji umożliwia wyłączenie podziału bez zmiany ustawień zmiennych).

Opcje modelu Apriori

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu, w przypadkach gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Zmienna unikalna. Określa zmienną używaną w celu unikalnej identyfikacji każdej obserwacji. Na przykład może to być zmienna identyfikacyjna, taka jak *CustomerID*. Produkt IBM SPSS Modeler nakłada ograniczenie, że zmienna kluczowa musi być liczbowa.

Uwaga: ta zmienna jest opcjonalna dla wszystkich węzłów Oracle za wyjątkiem Oracle Adaptive Bayes, Oracle O-Cluster i Oracle Apriori.

Automatyczne przygotowanie danych. (Dotyczy tylko 11g) Ta opcja włącza (stan domyślny) lub wyłącza tryb automatycznego przygotowywania danych produktu Oracle Data Mining. Jeśli to pole wyboru jest zaznaczone, wówczas ODM automatycznie przeprowadza transformacje danych wymagane przez algorytm. Więcej informacji zawiera dokumentacja *Koncepcje w programie Oracle Data Mining*.

Maksymalna długość reguły. Ustawia maksymalną liczbę warunków wstępnych dla reguły. Jest to wartość całkowita od 2 do 20. Jest to sposób na ograniczenie złożoności reguł. Jeśli reguły są zbyt złożone lub zbyt swoiste lub jeśli reguła zajmuje zbyt dużo czasu podczas uczenia, należy podjąć próbę zmniejszenia tego ustawienia.

Minimalna ufność. Ustawia minimalny poziom ufności, wartość od 0 do 1. Reguły o ufności niższej od podanego kryterium są odrzucane.

Minimalne wsparcie. Ustawia minimalny próg obsługi, wartość od 0 do 1. Algorytm Apriori wykrywa wzorce o częstotliwości powyżej minimalnego progu obsługi.

Oracle Minimum Description Length (MDL)

Algorytm Oracle Minimum Description Length (MDL) ułatwia rozpoznawanie atrybutów, które mają największy wpływ na atrybut przewidywany. Często wiedza o tym, które atrybuty mają największy wpływ, ułatwia zrozumienie i zarządzanie działalnością, a także pomaga upraszczać działania z zakresu modelowania. Ponadto te atrybuty mogą wskazywać typy danych, które użytkownik może dodać w celu doskonalenia modeli. Algorytm MDL może być używany na przykład w celu znajdowania atrybutów procesu, które są najbardziej istotne do przewidzenia jakości wyprodukowanej części, czynników związanych z odchodzeniem albo genów, które mogą być zaangażowane w leczenie konkretnej choroby.

Oracle MDL odrzuca zmienne wejściowe, które traktuje jako nieistotne w przewidywaniu zmiennej przewidywanej. Następnie na podstawie pozostałych zmiennych wejściowych algorytm buduje surowy model użytkowy, który jest powiązany z modelem Oracle i widoczny w produkcie Oracle Data Miner. Podczas przeglądania modelu w produkcie Oracle Data Miner wyświetlany jest wykres przedstawiający pozostałe pola wejściowe, którym przypisano rangi według ich istotności w predykcji zmiennej przewidywanej.

Ranga ujemna oznacza szum. Zmienne wejściowe, które mają rangę zero lub niższą, nie przyczyniają się do predykcji i prawdopodobnie powinny zostać usunięte z danych.

Aby wyświetlić wykres

1. Kliknij prawym przyciskiem myszy surowy model użytkowy w palecie Modele i wybierz opcję **Przełóżaj**.
2. W oknie modelu kliknij przycisk, aby uruchomić produkt Oracle Data Miner.
3. Nawiąż połączenie z produktem Oracle Data Miner. Więcej informacji można znaleźć w temacie “Oracle Data Miner” na stronie 46.
4. W panelu nawigatora Oracle Data Miner rozwiń pole **Modele**, a następnie **Znaczenie atrybutu**.
5. Wybierz odpowiedni model Oracle (będzie miał tę samą nazwę, co zmienna przewidywana określona w produkcie IBM SPSS Modeler). Jeśli nie wiesz, która nazwa jest poprawna, wybierz folder Attribute Importance i poszukaj modelu po dacie utworzenia.

Opcje modelu MDL

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu, w przypadkach gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Zmienna unikalna. Określa zmienną używaną w celu unikalnej identyfikacji każdej obserwacji. Na przykład może to być zmienna identyfikacyjna, taka jak *CustomerID*. Produkt IBM SPSS Modeler nakłada ograniczenie, że zmienna kluczowa musi być liczbowa.

Uwaga: ta zmienna jest opcjonalna dla wszystkich węzłów Oracle za wyjątkiem Oracle Adaptive Bayes, Oracle O-Cluster i Oracle Apriori.

Automatyczne przygotowanie danych. (Dotyczy tylko 11g) Ta opcja włącza (stan domyślny) lub wyłącza tryb automatycznego przygotowywania danych produktu Oracle Data Mining. Jeśli to pole wyboru jest zaznaczone, wówczas ODM automatycznie przeprowadza transformacje danych wymagane przez algorytm. Więcej informacji zawiera dokumentacja *Koncepcje w programie Oracle Data Mining*.

Oracle Attribute Importance (AI)

Celem znaczenia atrybutu jest znalezienie w zestawie danych takich atrybutów, które są powiązane z wynikiem, a także stopnia, w jakim wpływają na wynik końcowy. Węzeł Oracle Attribute Importance analizuje dane, znajduje wzorce i przewiduje rezultaty lub wyniki z powiązonym poziomem ufności.

Opcje modelu AI

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu, w przypadkach gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Użyj danych podzielonych na podzbiory. Jeśli zdefiniowano zmienną dzielącą na podzbiory, ta opcja umożliwia użycie podczas budowania modelu wyłącznie danych z podzbioru uczącego.

Automatyczne przygotowanie danych. (Dotyczy tylko 11g) Ta opcja włącza (stan domyślny) lub wyłącza tryb automatycznego przygotowywania danych produktu Oracle Data Mining. Jeśli to pole wyboru jest zaznaczone, wówczas ODM automatycznie przeprowadza transformacje danych wymagane przez algorytm. Więcej informacji zawiera dokumentacja *Koncepcje w programie Oracle Data Mining*.

Opcje doboru AI

Na karcie Opcje można określić domyślne ustawienia wybierania lub wykluczania zmiennych wejściowych w modelu użytkowym. Następnie można dodać model do strumienia, aby wybrać podzbiór zmiennych, jakie będą używane w kolejnych działaniach związanych z budowaniem modelu. Można również zastąpić te ustawienia, zaznaczając lub usuwając zaznaczenie dodatkowych zmiennych w przeglądarce modelu po wygenerowaniu modelu. Ustawienia domyślne umożliwiają jednak zastosowanie modelu użytkowego bez wprowadzania zmian, co może być przydatne w przypadku tworzenia skryptów.

Dostępne są następujące opcje:

Wszystkie zmienne z rangą. Wybiera zmienne na podstawie ich rangi: *important*, *marginal* lub *unimportant*. Można przeprowadzić edycję etykiety każdej rangi oraz wartości odcięcia użytych do przypisania rekordów do określonej rangi.

Określona liczba najważniejszych zmiennych. Wybiera pierwszych n zmiennych na podstawie ważności.

Ważność wyższa niż. Wybiera wszystkie zmienne, których ważność jest większa od określonej wartości.

Zmienna przewidywana jest zawsze zachowywana niezależnie od wyboru.

Model użytkowy AI — karta Model

Na karcie Model modelu użytkowego Oracle AI wyświetlana jest ranga i ważność wszystkich zmiennych wejściowych; można tu wybrać zmienne do filtrowania, używając pól wyboru w kolumnie po lewej stronie. Po uruchomieniu strumienia zachowane zostaną tylko zaznaczone zmienne razem z predykcją przewidywaną. Inne zmienne wejściowe są odrzucane. Wybór domyślny dokonywany jest na podstawie opcji określonych w węźle modelowania, jednak w razie potrzeby można zaznaczyć lub usunąć zaznaczenie dodatkowych zmiennych.

- Aby posortować listę według rangi, nazwy zmiennej, ważności lub innej wyświetlanej kolumny, należy kliknąć nagłówek kolumny. Alternatywnie wybierz żądany element z listy obok przycisku Sortuj według, a następnie użyj strzałek w górę i w dół, aby zmienić kolejność sortowania.
- Pasek narzędzi pozwala zaznaczyć lub usunąć zaznaczenie wszystkich zmiennych oraz uzyskać dostęp do okna dialogowego Zaznacz zmienne, co pozwoli zaznaczyć zmienne według rangi lub ważności. Można także nacisnąć klawisze Shift i Ctrl, klikając jednocześnie zmienne, aby rozszerzyć wybór.
- Wartości graniczne dla rangowania zmiennych wejściowych, takie jak ważne, brzegowe lub nieważne, są wyświetlane w legendzie pod tabelą. Wartości te są określane w węźle modelowania.

Zarządzanie modelami Oracle

Modele Oracle są dodawane do palety Modele tak samo, jak inne modele IBM SPSS Modeler, a ponadto mogą być używane w bardzo podobny sposób. Istnieje jednak kilka istotnych różnic, ponieważ każdy model Oracle utworzony w produkcie IBM SPSS Modeler w rzeczywistości odwołuje się do modelu zapisanego na serwerze bazy danych.

Karta serwera modelu użytkowego Oracle

Budowanie modelu ODM za pośrednictwem produktu IBM SPSS Modeler powoduje utworzenie modelu w produkcie IBM SPSS Modeler oraz utworzenie lub zastąpienie modelu w bazie danych Oracle. Model IBM SPSS Modeler odwołuje się do zawartości modelu bazy danych zapisanego na serwerze bazy danych. Produkt IBM SPSS Modeler może przeprowadzić sprawdzanie spójności poprzez zapisanie identycznego wygenerowanego **klucza tekstowego modelu** zarówno w modelu IBM SPSS Modeler, jak i w modelu Oracle.

Klucz tekstowy dla każdego modelu Oracle jest wyświetlany w kolumnie *Informacje o modelu* w oknie dialogowym Wymień modele. Klucz tekstowy dla modelu IBM SPSS Modeler jest wyświetlany jako **klucz modelu** na karcie Serwer modelu IBM SPSS Modeler (po umieszczeniu w strumieniu).

Przycisku Sprawdź, który znajduje się na karcie Serwer modelu użytkowego, można użyć, aby sprawdzić, czy klucze modelu w modelu IBM SPSS Modeler i modelu Oracle są zgodne. Jeśli w Oracle nie można znaleźć żadnego modelu o

tej samej nazwie lub jeśli klucze modelu są niezgodne, wówczas oznacza to, że model Oracle został usunięty lub ponownie zbudowany po zbudowaniu modelu IBM SPSS Modeler.

Karta podsumowania modelu użytkowego Oracle

Karta Podsumowanie modelu użytkowego zawiera informacje na temat samego modelu (*Analiza*), zmiennych użytych w modelu (*Zmienne*), ustawień użytych podczas budowania modelu (*Ustawienia budowania*) i uczenia modelu (*Podsumowanie uczenia*).

Podczas przeglądania węzła po raz pierwszy karta Podsumowanie jest zwinięta. Aby zobaczyć wyniki będące przedmiotem zainteresowania, należy użyć rozszerzanego elementu sterującego po lewej stronie pozycji, aby ją rozwinąć, lub kliknąć przycisk **Rozwiń wszystko**, aby wyświetlić wszystkie wyniki. W celu ukrycia wyników po zakończeniu ich przeglądania należy użyć rozszerzanego elementu sterującego, aby zwinąć konkretne wyniki, jakie mają zostać ukryte, lub kliknąć przycisk **Zwiń wszystko**, aby zwinąć wszystkie wyniki.

Analiza. Wyświetla informacje na temat konkretnego modelu. Jeśli wykonano węzeł analizy dołączony do modelu użytkowego, wówczas informacje z tej analizy również pojawiają się w tej sekcji.

Zmienne. Na liście znajdują się zmienne użyte jako zmienne przewidywane i wejściowe podczas budowania modelu.

Ustawienia budowania. Zawiera informacje na temat ustawień użytych podczas budowania modelu.

Podsumowanie uczenia. Przedstawia typ modelu, strumień użyty do jego utworzenia, użytkownika, który go utworzył, informację, kiedy został utworzony, oraz czas, jaki był potrzebny do zbudowania modelu.

Karta ustawień modelu użytkowego Oracle

Karta Ustawienia w modelu użytkowym umożliwia zastąpienie ustawienia niektórych opcji w węźle modelowania na potrzeby oceniania.

Oracle Decision Tree

Stosuj koszty błędnej klasyfikacji. Określa, czy koszty błędnej klasyfikacji będą stosowane w modelu Oracle Decision Tree. Więcej informacji można znaleźć w temacie “Koszty błędnej klasyfikacji” na stronie 30.

Identyfikator reguły. Jeśli ta opcja jest zaznaczona, dodaje kolumnę identyfikatora reguły do modelu Oracle Decision Tree. Identyfikator reguły identyfikuje węzeł w drzewie, w którym wykonywany jest konkretny podział.

Oracle NMF

Pokaż wszystkie predyktory. Jeśli to pole wyboru jest zaznaczone, wówczas w modelu Oracle NMF wyświetlane są identyfikator i ufnosc predyktora dla wszystkich predyktorów, a nie tylko wartości dla najlepszego predyktora.

Lista modeli Oracle

Przycisk Wymień modele Oracle Data Mining umożliwia wyświetlenie okna dialogowego, które zawiera listę istniejących modeli bazy danych i w którym te modele można usuwać. To okno dialogowe może być uruchamiane z okna dialogowego Aplikacje pomocnicze, a także z okien dialogowych do budowania, przeglądania i okna dialogowego Zastosuj dla węzłów dotyczących ODM.

Dla każdego modelu wyświetlane są następujące informacje:

- **Nazwa modelu.** Nazwa modelu, która jest używana w celu sortowania listy
- **Informacje o modelu.** Informacje o kluczu modelu, które obejmują datę/godzinę zbudowania, a także nazwę kolumny zmiennych przewidywanych
- **Typ modelu.** Nazwa algorytmu, który zbudował ten model

Oracle Data Miner

Oracle Data Miner to interfejs użytkownika produktu Oracle Data Mining (ODM), który zastępuje poprzedni interfejs użytkownika IBM SPSS Modeler dla produktu ODM. Oracle Data Miner jest przeznaczony do zwiększania częstotliwości stosowania poprawnych algorytmów ODM. Te cele są realizowane na kilka sposobów:

- Użytkownicy potrzebują więcej wsparcia w stosowaniu metodologii, która dotyczy przygotowania danych i wyboru algorytmów. Oracle Data Miner spełnia tę potrzebę, ponieważ udostępnia działania eksploracji danych, dzięki którym kieruje użytkowników przez odpowiednie metodologie.
- Oracle Data Miner zawiera poprawione i rozszerzone heurystyki w kreatorach budowania modeli i transformacji, co zmniejsza prawdopodobieństwo błędów podczas określania ustawień modelu i transformacji.

Definiowanie połączenia w interfejsie Oracle Data Miner

1. Interfejs Oracle Data Miner można uruchomić z wszystkich węzłów budowania i węzłów stosowania Oracle, a także z okien dialogowych wyników, korzystając z przycisku **Uruchom Oracle Data Miner**.



Rysunek 2. Przycisk Uruchom Oracle Data Miner

2. Okno dialogowe **Edytuj połączenie** w interfejsie Oracle Data Miner jest prezentowane użytkownikowi przed uruchomieniem aplikacji zewnętrznej Oracle Data Miner (pod warunkiem że opcja aplikacji pomocniczej jest poprawnie zdefiniowana).
Uwaga: to okno dialogowe jest wyświetlane wyłącznie przy braku nazwy zdefiniowanego połączenia.
 - Udostępnij nazwę połączenia Data Miner i wprowadź informacje o odpowiednim serwerze Oracle 10gR1 lub 10gR2. Serwer Oracle powinien być tym samym serwerem, który jest określony w produkcie IBM SPSS Modeler.
3. Okno dialogowe **Wybierz połączenie** w interfejsie Oracle Data Miner zawiera opcje, które umożliwiają określenie nazwy używanego połączenia, co zostało zdefiniowane w kroku powyżej.

Więcej informacji na temat wymagań, instalacji i zastosowań interfejsu Oracle Data Miner zawiera strona poświęcona Oracle Data Miner w serwisie WWW Oracle.

Przygotowywanie danych

W przypadku stosowania podczas modelowania algorytmów Naive Bayes, Adaptive Bayes i SVM, które są dostępne wśród algorytmów Oracle Data Mining, użyteczne są dwa rodzaje przygotowań danych:

- **Kategoryzacja**, czyli przekształcenie zmiennych będących ciągłymi przedziałami liczbowymi na kategorie dla algorytmów, które nie mogą akceptować danych ciągłych.
- **Normalizacja**, czyli transformacje stosowane względem zakresów liczbowych, dzięki którym zakresy mają podobne średnie i odchylenia standardowe.

Kategoryzacja

Węzeł kategoryzacji w produkcie IBM SPSS Modeler oferuje szereg technik przeznaczonych do wykonywania operacji kategoryzacji. Zdefiniowana jest operacja kategoryzacji, którą można zastosować względem co najmniej jednej zmiennej. Wykonanie operacji kategoryzacji względem zestawu danych powoduje utworzenie progów i umożliwia utworzenie węzła wyliczeń w produkcie IBM SPSS Modeler. Operację wyliczeń można przekształcić na kod SQL i zastosować przed budowaniem i oceną modelu. Takie podejście powoduje powstanie zależności między modelem a węzłem wyliczeń, który realizuje kategoryzację, ale pozwala na ponowne użycie specyfikacji kategoryzacji w wielu zadaniach modelowania.

Normalizacja

Zmienne ciągłe (zakres liczbowy), które są używane jako dane wejściowe dla modeli SVM, powinny być znormalizowane przed budowaniem modelu. W przypadku modeli regresji normalizacja musi również zostać odwrócona w celu rekonstrukcji oceny z danych wyjściowych modelu. Ustawienia modelu SVM umożliwiają wybór opcji **Statystyki z, Min.-Maks.** lub **Brak**. Współczynniki normalizacji są konstruowane przez Oracle jako etap procesu budowania modelu, a następnie współczynniki są wprowadzane do produktu IBM SPSS Modeler i zapisywane w modelu. Po zastosowaniu współczynniki są przekształcane na wyrażenia wyliczeń w produkcie IBM SPSS Modeler i używane w celu przygotowania danych do oceny przed przekazaniem danych do modelu. W tym przypadku normalizacja jest ściśle powiązana z zadaniem modelowania.

Przykłady dotyczące Oracle Data Mining

Przedstawiono szereg strumieni, które demonstrują użycie produktu ODM z produktem IBM SPSS Modeler. Te strumienie można znaleźć w folderze instalacyjnym produktu IBM SPSS Modeler w katalogu `\Demos\Database_Modelling\Oracle Data Mining\`.

Uwaga: dostęp do folderu Demos można uzyskać z grupy programu IBM SPSS Modeler w menu Start systemu Windows.

Strumienie przedstawione w poniższej tabeli mogą być używane razem w kolejności jako przykład procesu eksploracji bazy danych, przy użyciu algorytmu Support Vector Machine (SVM), który jest udostępniany razem z produktem Oracle Data Mining:

Tabela 4. Eksploracja danych — przykładowe strumienie

Strumień	Opis
<code>1_upload_data.str</code>	Służy do kasowania i wczytywania danych z pliku płaskiego do bazy danych.
<code>2_explore_data.str</code>	Udostępnia przykład eksploracji danych z użyciem produktu IBM SPSS Modeler
<code>3_build_model.str</code>	Buduje model z użyciem algorytmu rodzimego dla bazy danych.
<code>4_evaluate_model.str</code>	Używany jako przykład oceny modelu za pomocą produktu IBM SPSS Modeler
<code>5_deploy_model.str</code>	Wdraża model na potrzeby oceniania w bazie danych.

Uwaga: w celu uruchomienia tego przykładu strumienie muszą być wykonywane kolejno. Ponadto węzły źródłowe i węzły modelowania w każdym strumieniu muszą być aktualizowane w taki sposób, aby odwoływały się do poprawnego źródła danych dla bazy danych przeznaczonej do użycia.

Zestaw danych używany w strumieniach przykładowych dotyczy aplikacji kart kredytowych i prezentuje problem klasyfikacji z mieszaniną predyktorów jakościowych i ciągłych. Więcej informacji na temat tego zestawu danych zawiera plik `crx.names`, który znajduje się w tym samym folderze, w którym są przykładowe strumienie.

Ten zestaw danych jest dostępny z repozytorium UCI Machine Learning Repository na stronie <ftp://ftp.ics.uci.edu/pub/machine-learning-databases/credit-screening/>.

Przykładowy strumień: Wczytywanie danych

Pierwszy przykładowy strumień o nazwie `1_upload_data.str` jest używany do kasowania i wczytywania danych z pliku płaskiego do programu Oracle.

Rozwiązanie Oracle Data Mining wymaga zmiennej o unikalnym identyfikatorze, dlatego ten strumień początkowy używa węzła wyliczeń, aby dodać nową zmienną do zestawu danych o nazwie `ID` i unikalnych wartościach 1,2,3, przy użyciu funkcji IBM SPSS Modeler `@INDEX`.

Węzeł wypełniania jest używany do obsługi wartości brakujących i zastępuje puste pola odczytywane z pliku tekstowego *crx.data* wartościami *NULL*.

Przykładowy strumień: Eksploracja danych

Drugi przykładowy strumień — *2_explore_data.str* — jest używany do zaprezentowania użycia węzła Audyt danych w celu uzyskania ogólnego przeglądu danych, w tym statystyk i wykresów podsumowujących.

Dwukrotne kliknięcie wykresu w raporcie z audytu danych generuje bardziej szczegółowy wykres, który pozwala na głębszą eksplorację pod kątem konkretnej zmiennej.

Przykładowy strumień: Budowa modelu

Trzeci przykładowy strumień — *3_build_model.str* — ilustruje budowanie modelu w produkcie IBM SPSS Modeler. Kliknij dwukrotnie węzeł źródła bazy danych (oznaczony jako CREDIT), aby określić źródło danych. Aby określić ustawienia budowania, kliknij dwukrotnie węzeł budowania (początkowo oznaczony jako CLASS, co po określeniu źródła danych ulega zmianie na FIELD16).

Na karcie Model w oknie dialogowym:

1. Upewnij się, że jako zmienna unikalna wybrana jest wartość **ID**.
2. Upewnij się, że wartość **Liniowy** jest wybrana jako funkcja algorytmu domyślnego, a wartość **Statystyki z** jest metodą normalizacji.

Przykładowy strumień: Ocena modelu

Czwarty przykładowy strumień — *4_evaluate_model.str* — ilustruje zalety korzystania z produktu IBM SPSS Modeler na potrzeby modelowania w bazie danych. Po wykonaniu modelu można go dodać z powrotem do strumienia danych i ocenić model, używając kilku narzędzi oferowanych w produkcie IBM SPSS Modeler.

Wyświetlanie wyników modelowania

Dołącz węzeł tabeli do modelu użytkowego, aby zapoznać się z wynikami. Zmienna **\$O-field16** przedstawia wartość przewidywaną dla *field16* w każdej obserwacji, a zmienna **\$OC-field16** przedstawia współczynnik ufności dla tej predykcji.

Ocena wyników modelu

Węzła analizy można użyć, aby utworzyć macierz zbieżności przedstawiającą wzorzec dopasowań między poszczególnymi zmiennymi przewidywanymi a odpowiadającymi im zmiennymi zależnymi. Uruchom węzeł analizy, aby zobaczyć wyniki.

Węzła ewaluacji można użyć, aby utworzyć wykres korzyści przeznaczony do przedstawienia postępów modelu w zakresie dokładności. Uruchom węzeł ewaluacji, aby zobaczyć wyniki.

Przykładowy strumień: Wdrożenie modelu

Gdy dokładność modelu jest zadowalająca, można go wdrożyć, aby używać go z aplikacjami zewnętrznymi albo opublikować z powrotem do bazy danych. W ostatnim strumieniu przykładowym o nazwie *5_deploy_model.str* dane są odczytywane z tabeli CREDITDATA, a następnie oceniane i publikowane do tabeli CREDITSCORES przy użyciu węzła publikacji o nazwie *deploy solution*.

Rozdział 5. Modelowanie w bazie za pomocą produktu IBM InfoSphere Warehouse

IBM InfoSphere Warehouse i produkt IBM SPSS Modeler

Rozwiązanie IBM InfoSphere Warehouse (ISW) udostępnia rodzinę algorytmów eksploracji danych osadzonych w systemie RDBMS IBM DB2. Produkt IBM SPSS Modeler udostępnia węzły, które obsługują integrację następujących algorytmów IBM:

- Drzewa decyzyjne
- Reguły asocjacyjne
- Grupowanie demograficzne
- Grupowanie Kohonena
- Reguły sekwencyjne
- Regresja z transformacją
- regresja liniowa
- Regresja wielomianowa
- Naive Bayes
- Regresja logistyczna
- Szereg czasowy

Więcej informacji na temat tych algorytmów znajduje się w dokumentacji dołączonej do instalacji produktu IBM InfoSphere Warehouse.

Wymagania dotyczące integracji z produktem IBM InfoSphere Warehouse

Następujące warunki stanowią warunki wstępne do przeprowadzenia modelowania w bazie danych z użyciem produktu InfoSphere Warehouse Data Mining. W celu upewnienia się, że te warunki są spełnione, konieczne mogą być konsultacje z administratorem bazy danych.

- Produkt IBM SPSS Modeler działający względem instalacji IBM SPSS Modeler Server w systemie Windows lub UNIX.
- IBM DB2 Data Warehouse Edition, wersja 9.1
lub
- IBM InfoSphere Warehouse, wersja 9.5 Enterprise Edition
- Źródło danych ODBC w celu nawiązywania połączeń z DB2 zgodnie z opisem poniżej.

Uwaga: Modelowanie w bazie danych i optymalizacja SQL wymagają włączenia na komputerze z programem IBM SPSS Modeler możliwości połączenia z serwerem IBM SPSS Modeler Server. Po włączeniu tej opcji można uzyskać dostęp do algorytmów baz danych, wstawić SQL do kolejki bezpośrednio z programu IBM SPSS Modeler i uzyskać dostęp do programu IBM SPSS Modeler Server. W celu sprawdzenia bieżącego statusu licencji należy wybrać z menu programu IBM SPSS Modeler następujące opcje.

Pomoc > Informacje o programie > Dodatkowe szczegóły

Po włączeniu możliwości połączenia na karcie Status licencji widoczna jest opcja **Aktywacja serwera**.

Włączanie integracji z produktem IBM InfoSphere Warehouse

Aby włączyć integrację produktu IBM SPSS Modeler z produktem IBM InfoSphere Warehouse (ISW) Data Mining, należy skonfigurować produkt ISW, utworzyć źródło ODBC, włączyć integrację w oknie dialogowym Aplikacje pomocnicze w produkcie IBM SPSS Modeler, a także włączyć generowanie i optymalizację kodu SQL.

Konfigurowanie produktu ISW

Aby zainstalować i skonfigurować produkt ISW, należy postępować zgodnie z instrukcjami zawartymi w przewodniku instalacji *InfoSphere Warehouse*.

Tworzenie źródła ODBC dla produktu ISW

Aby umożliwić połączenie produktu ISW i produktu IBM SPSS Modeler, należy utworzyć systemową nazwę źródła danych (DSN) ODBC.

Przed utworzeniem DSN wymagana jest podstawowa znajomość źródeł danych i sterowników ODBC, a także obsługa bazy danych w produkcie IBM SPSS Modeler.

Jeśli produkt IBM SPSS Modeler Server i produkt IBM InfoSphere Warehouse Data Mining działają na różnych hostach, utwórz tę samą nazwę DSN ODBC na każdym hoście. Upewnij się, że na każdym hoście nazwa DSN jest taka sama.

1. Zainstaluj sterowniki ODBC. Są dostępne na dysku instalacyjnym produktu IBM SPSS Data Access Pack dostarczonym z tym wydaniem. Uruchom plik *setup.exe*, aby uruchomić instalatora i wybrać wszystkie odpowiednie sterowniki. W celu zainstalowania sterowników postępuj zgodnie z instrukcjami wyświetlanymi na ekranie.
 - a. Utwórz DSN.

Uwaga: kolejność opcji w menu jest zależna od wersji systemu Windows.

 - **Windows XP.** Z menu Start wybierz opcję **Panel sterowania**. Kliknij dwukrotnie opcję **Narzędzia administracyjne**, a następnie kliknij dwukrotnie opcję **Źródła danych (ODBC)**.
 - **Windows Vista.** Z menu Start wybierz opcję **Panel sterowania**, a następnie **Konserwacja systemu**. Kliknij dwukrotnie pozycję **Narzędzia administracyjne**, wybierz pozycję **Źródła danych (ODBC)**, a następnie kliknij opcję **Otwórz**.
 - **Windows 7.** Z menu Start wybierz pozycję **Panel sterowania**, następnie **System i zabezpieczenia**, a następnie **Narzędzia administracyjne**. Wybierz pozycję **Źródła danych (ODBC)**, a następnie kliknij opcję **Otwórz**.
 - b. Kliknij kartę **Systemowe źródło danych DSN** i kliknij opcję **Dodaj**.
2. Wybierz sterownik **SPSS OEM 6.0 DB2 Wire Protocol**.
3. Kliknij przycisk **Zakończ**.
4. W oknie dialogowym ODBC DB2 Wire Protocol Driver Setup (Konfiguracja sterownika ODBC DB2 Wire Protocol Driver):
 - Określ nazwę źródła danych.
 - Jako adres IP podaj nazwę hosta serwera, na którym znajduje się system RDBMS DB2.
 - Zaakceptuj domyślny port TCP (50000).
 - Określ nazwę bazy danych, z którą będziesz nawiązywać połączenie.
5. Kliknij opcję **Test Connection** (Testuj połączenie).
6. W oknie dialogowym Logon to DB2 Wire Protocol (Logowanie do protokołu DB2 Wire Protocol) wprowadź nazwę i hasło użytkownika udostępnione przez administratora bazy danych, a następnie kliknij przycisk **OK**. Zostanie wyświetlony komunikat **Connection established!** (Połączenie nawiązane!).

IBM DB2 ODBC DRIVER. Jeśli Twoim sterownikiem ODBC jest IBM DB2 ODBC DRIVER, wykonaj poniższe kroki, aby utworzyć DSN ODBC:

7. W obszarze administratora źródła danych ODBC kliknij kartę **System DSN** (Systemowe źródło danych DSN), a następnie kliknij opcję **Add** (Dodaj).
8. Wybierz **IBM DB2 ODBC DRIVER** i kliknij opcję **Finish** (Zakończ).
9. W oknie dodawania IBM DB2 ODBC DRIVER wprowadź nazwę źródła danych, a następnie kliknij opcję **Dodaj** dla aliasu bazy danych.
10. W oknie ustawień CLI/ODBC — <Data source name> (Nazwa źródła danych), na karcie źródła danych wprowadź identyfikator użytkownika i hasło udostępnione przez administratora bazy danych, a następnie kliknij kartę **TCP/IP**.
11. Na karcie TCP/IP wprowadź:
 - Nazwę bazy danych, z którą będziesz nawiązywać połączenie.
 - Alias bazy danych (nie więcej niż osiem znaków).
 - Nazwę hosta serwera bazy danych, z którą będziesz nawiązywać połączenie.
 - Numer portu dla połączenia.
12. Kliknij kartę **Security Options** (Opcje zabezpieczeń) i wybierz opcję **Specify the security options (Optional)** (Określ opcje zabezpieczeń (opcjonalnie)), a następnie zaakceptuj ustawienie domyślne (**Use authentication value in server's DBM Configuration** (Użyj wartości uwierzytelniania w konfiguracji DBM serwera)).
13. Kliknij kartę **Data Source** (Źródło danych) i kliknij opcję **Connect** (Połącz).

Zostanie wyświetlony komunikat **Connection tested successfully** (Połączenie pomyślnie przetestowane).

Konfigurowanie ODBC na potrzeby informacji zwrotnych (opcjonalnie)

Aby uzyskiwać informacje zwrotne z produktu IBM InfoSphere Warehouse Data Mining podczas budowania modelu i zezwolić produktowi IBM SPSS Modeler na anulowanie budowania modelu, należy wykonać poniższe czynności w celu skonfigurowania źródła danych ODBC, które zostało utworzone w poprzedniej sekcji. Zwróć uwagę na to, że ten etap konfiguracji umożliwia produktowi IBM SPSS Modeler odczyt danych DB2, które nie mogą zostać zatwierdzone do bazy danych przez transakcje wykonywane współbieżnie. W przypadku wątpliwości związanych ze skutkami tej zmiany należy się skontaktować z administratorem bazy danych.

SPSS OEM 6.0 DB2 Wire Protocol Driver. W przypadku sterownika połączenia ODBC należy wykonać poniższe czynności:

1. Uruchom administratora źródła danych ODBC, wybierz źródło danych utworzone w poprzedniej sekcji i kliknij przycisk **Configure** (Konfiguruj).
 2. W oknie dialogowym ODBC DB2 Wire Protocol Driver Setup (Konfiguracja sterownika ODBC DB2 Wire Protocol Driver) kliknij kartę **Advanced** (Zaawansowane).
 3. Ustaw domyślny poziom odseparowania na **0-READ UNCOMMITTED**, a następnie kliknij przycisk **OK**.
- IBM DB2 ODBC Driver.** W przypadku sterownika IBM DB2 należy wykonać poniższe czynności:
4. Uruchom administratora źródła danych ODBC, wybierz źródło danych utworzone w poprzedniej sekcji i kliknij przycisk **Configure** (Konfiguruj).
 5. W oknie dialogowych ustawień CLI/ODBC kliknij kartę **Advanced Settings** (Ustawienia zaawansowane), a następnie kliknij przycisk **Add** (Dodaj).
 6. W oknie dialogowym dodawania parametru CLI/ODBC wybierz parametr **TXNISOLATION** i kliknij przycisk **OK**.
 7. W oknie dialogowym poziomu odseparowania wybierz opcję **Read Uncommitted** (Odczyt niepotwierdzonych) i kliknij przycisk **OK**.
 8. W oknie dialogowym ustawień CLI/ODBC kliknij przycisk **OK**, aby zakończyć konfigurację.

Zwróć uwagę na to, że informacje zwrotne zgłaszane przez produkt IBM InfoSphere Warehouse Data Mining są wyświetlane w następującym formacie:

```
<ITERATIONNO> / <PROGRESS> / <KERNELPHASE>
```

gdzie:

- <ITERATIONNO> wskazuje numer aktualnego wykonania względem danych, począwszy od 1.
- <PROGRESS> wskazuje postęp bieżącej iteracji w postaci liczby od 0.0 do 1.0.
- <KERNELPHASE> opisuje bieżącą fazę algorytmu eksploracji.

Włączanie integracji z produktem IBM InfoSphere Warehouse Data Mining Integration w produkcie IBM SPSS Modeler

Aby umożliwić produktowi IBM SPSS Modeler korzystanie z bazy danych DB2 z produktem IBM InfoSphere Warehouse Data Mining, należy najpierw udostępnić pewne specyfikacje w oknie dialogowym Aplikacje pomocnicze.

1. Z menu programu IBM SPSS Modeler wybierz:

Narzędzia > Opcje > Aplikacje pomocnicze

2. Kliknij kartę **IBM InfoSphere Warehouse**.

Włącz integrację z InfoSphere Warehouse Data Mining. Włącza paletę Modelowanie w bazie (jeśli nie jest jeszcze wyświetlana) u dołu okna IBM SPSS Modeler i dodaje węzły przeznaczone dla algorytmów ISW Data Mining.

Połączenie z bazą DB2. Określa domyślne źródło danych ODBC DB2 używane na potrzeby budowania i zapisywania modeli. To ustawienie może zostać przesłonięte w pojedynczych węzłach budowania modeli i wygenerowanych węzłach modeli. Kliknij przycisk z wielokropkiem (...), aby wybrać źródło danych.

Połączenie z bazą danych używane na potrzeby modelowania może być takie samo, jak połączenie używane w celu uzyskiwania dostępu do danych. Na przykład może istnieć strumień, który uzyskuje dostęp do danych z jednej z baz danych DB2, pobiera dane do produktu IBM SPSS Modeler w celu kasowania lub w celu przeprowadzenia innych manipulacji, a następnie przesyła te dane do innej bazy danych DB2 na potrzeby modelowania. Alternatywnie oryginalne dane mogą istnieć w pliku płaskim lub innym źródle (innym niż DB2), ale w takim przypadku na potrzeby modelowania konieczne będzie przesłanie ich do bazy danych DB2. We wszystkich przypadkach w razie potrzeby dane będą automatycznie wprowadzane do tabeli tymczasowej utworzonej w bazie danych używanej na potrzeby modelowania.

Ostrzegaj przed nadpisaniem modelu InfoSphere Warehouse Data Mining. Wybierz tę opcję, aby upewnić się, że modele zapisane w bazie danych nie będą zastępowane przez produkt IBM SPSS Modeler bez ostrzeżenia.

Przedstaw modele InfoSphere Warehouse Data Mining. Ta opcja umożliwia wyświetlenie listy i usuwanie modeli zapisanych w DB2. Więcej informacji można znaleźć w temacie “Lista modeli bazy danych” na stronie 54.

Pozwól na uruchamianie InfoSphere Warehouse Data Mining Visualization. Jeśli zainstalowany został moduł Visualization, należy włączyć go w tym miejscu, aby mógł z niego korzystać produkt IBM SPSS Modeler.

Ścieżka do uruchomienia aplikacji Visualization. Położenie pliku wykonywalnego modułu Visualization (jeśli jest zainstalowany), na przykład *C:\Program Files\IBM\ISWarehouse\Im\IMVisualization\bin\imvisualizer.exe*.

Katalog wtyczki wizualizacji szeregów czasowych. Położenie wtyczki flash wizualizacji szeregów czasowych (jeśli jest zainstalowana), na przykład *C:\Program Files\IBM\ISWShared\plugins\com.ibm.datatools.datamining.imvisualization.flash_2.2.1.v20091111_0915*.

Aktywuj opcje InfoSphere Warehouse Data Mining. Dla niektórych modeli można określić limit wykorzystania pamięci w ramach algorytmu eksploracji w bazie danych, a także określić dowolne opcje w wierszu komend. Limit pamięci umożliwia kontrolowanie wykorzystania pamięci i określenie wartości dla opcji zaawansowanej *-buf*. Inne opcje zaawansowane można określić tutaj w postaci wiersza komend, a następnie one są przekazywane do produktu IBM InfoSphere Warehouse Data Mining. Więcej informacji można znaleźć w temacie “Opcje zaawansowane” na stronie 56.

Sprawdź wersję systemu InfoSphere Warehouse. Sprawdza używaną wersję produktu IBM InfoSphere Warehouse i zgłasza błąd przy próbie użycia funkcji eksploracji danych, która nie jest obsługiwana w używanej wersji.

Włączenie opcji generowania i optymalizacji kodu SQL

1. Z menu programu IBM SPSS Modeler wybierz:
Narzędzia > Właściwości strumienia > Opcje
2. Kliknij opcję **Optymalizacja** w panelu nawigacji.
3. Upewnij się, że włączona jest opcja **Generuj kod SQL kierowany do bazy**. To ustawienie jest niezbędne, ponieważ zapewni poprawne działanie modelowania w bazie danych.
4. Wybierz opcje **Optymalizuj operacje generujące kod SQL** i **Optymalizuj inne wykonywane operacje** (nie jest to ściśle wymagane, ale zdecydowanie zalecane w celu poprawy wydajności).

Budowanie modelu za pomocą produktu IBM InfoSphere Warehouse Data Mining

W przypadku budowania modelu w IBM InfoSphere Warehouse Data Mining wymagane jest, aby uczący zbiór danych znajdował się w tabeli lub widoku w bazie danych DB2. Jeśli dane nie znajdują się w DB2 lub muszą być przetwarzane w produkcie IBM SPSS Modeler w ramach procesów przygotowywania danych, które mogą być wykonywane w DB2, wówczas dane są automatycznie wczytywane do tabeli tymczasowej w DB2 przed budowaniem modelu.

Ocenianie i wdrażanie modelu

Ocenianie modelu zawsze odbywa się w DB2 i zawsze jest wykonywane przez produkt IBM InfoSphere Warehouse Data Mining. Jeśli dane pochodzą z produktu IBM SPSS Modeler lub muszą być w nim przygotowywane, wówczas konieczne może być wczytanie zestawu danych do tabeli tymczasowej. W przypadku modelu drzewa decyzyjnego, modelu regresji i modelu grupowania w produkcie IBM SPSS Modeler zwykle zwracana jest tylko pojedyncza predykcja i powiązane prawdopodobieństwo lub współczynnik ufności. Ponadto dostępna dla użytkownika opcja, dzięki której możliwe jest wyświetlenie współczynników ufności dla każdego możliwego wyniku (co jest podobne do opcji dostępnej w regresji logistycznej) to opcja czasu oceny dostępna na karcie ustawień modelu użytkowego (pole wyboru **Dołącz ufności dla wszystkich klas**). W przypadku modeli asocjacyjnych i modeli sekwencji w produkcie IBM SPSS Modeler zwracanych jest kilka wartości. Produkt IBM SPSS Modeler może oceniać modele IBM InfoSphere Warehouse Data Mining ze strumieni opublikowanych w celu wykonania przez produkt IBM SPSS Modeler Solution Publisher.

W poniższej tabeli przedstawiono zmienne generowane przez modele oceniania.

Tabela 5. Zmienne oceniania modeli

Typ modelu	Kolumny ocen	Znaczenie
Drzewa decyzyjne	\$I-zmienna	Najlepsza predykcja dla <i>zmiennej</i> .
	\$IC-zmienna	Ufność najlepszej predykcji dla <i>zmiennej</i> .
	\$IC-wartość1, ..., \$IC-wartośćN	(opcjonalnie) Ufność każdej z możliwych <i>N</i> wartości dla <i>field</i> .
Regresja	\$I-zmienna	Najlepsza predykcja dla <i>zmiennej</i> .
	\$IC-zmienna	Ufność najlepszej predykcji dla <i>zmiennej</i> .
Grupowanie	\$I-nazwa_modelu	Najlepsze przypisanie skupień dla rekordu wejściowego.
	\$IC-nazwa_modelu	Ufność najlepszego przypisania skupień dla rekordu wejściowego.
Związek	\$I-nazwa_modelu	Identyfikator reguły dopasowania.
	\$IH-nazwa_modelu	Element nagłówka.
	\$IHN-nazwa_modelu	Nazwa elementu nagłówka.

Tabela 5. Zmienne oceniania modeli (kontynuacja)

Typ modelu	Kolumny ocen	Znaczenie
	\$IS-nazwa_modelu	Wartość pomocnicza reguły dopasowania.
	\$IC-nazwa_modelu	Wartość ufności reguły dopasowania.
	\$IL-nazwa_modelu	Wartość wzrostu reguły dopasowania.
	\$IMB-nazwa_modelu	Liczba dopasowanych elementów treści lub zbiorów elementów treści (ponieważ wszystkie elementy treści lub zbiory elementów treści muszą być zgodne z tą liczbą, jest ona równa liczbie elementów treści lub zbiorów elementów treści).
Sekwencje	\$I-nazwa_modelu	Identyfikator reguły dopasowania
	\$IH-nazwa_modelu	Zbiór elementu nagłówka reguły dopasowania
	\$IHN-nazwa_modelu	Nazwy elementów w zbiorze elementów nagłówka reguły dopasowania
	\$IS-nazwa_modelu	Wartość pomocnicza reguły dopasowania
	\$IC-nazwa_modelu	Wartość ufności reguły dopasowania
	\$IL-nazwa_modelu	Wartość wzrostu reguły dopasowania
	\$IMB-nazwa_modelu	Liczba dopasowanych elementów treści lub zbiorów elementów treści (wszystkie elementy treści lub zbiory elementów treści muszą być zgodne z tą liczbą, jest ona równa liczbie elementów treści lub zbiorów elementów treści).
Naive Bayes	\$I-zmienna	Najlepsza predykcja dla <i>zmiennej</i> .
	\$IC-zmienna	Ufność najlepszej predykcji dla <i>zmiennej</i> .
Regresja logistyczna	\$I-zmienna	Najlepsza predykcja dla <i>zmiennej</i> .
	\$IC-zmienna	Ufność najlepszej predykcji dla <i>zmiennej</i> .

Zarządzanie modelami DB2

Budowanie modelu IBM InfoSphere Warehouse Data Mining za pośrednictwem produktu IBM SPSS Modeler powoduje utworzenie modelu w produkcie IBM SPSS Modeler oraz utworzenie lub zastąpienie modelu w bazie danych DB2. Model tego rodzaju w produkcie IBM SPSS Modeler odwołuje się do zawartości modelu bazy danych zapisanego na serwerze bazy danych. Produkt IBM SPSS Modeler może przeprowadzić sprawdzanie spójności poprzez zapisanie identycznego wygenerowanego klucza tekstowego zarówno w modelu IBM SPSS Modeler, jak i w modelu DB2.

Klucz tekstowy dla każdego modelu DB2 jest wyświetlany w kolumnie *Informacje o modelu* w oknie dialogowym Wymień modele bazy danych. Klucz tekstowy dla modelu IBM SPSS Modeler jest wyświetlany jako klucz modelu na karcie Serwer modelu IBM SPSS Modeler (po umieszczeniu w strumieniu).

Przycisku Sprawdź można użyć, aby sprawdzić, czy klucze modelu w modelu IBM SPSS Modeler i modelu DB2 są zgodne. Jeśli w DB2 nie można znaleźć żadnego modelu o tej samej nazwie lub jeśli klucze modelu są niezgodne, wówczas oznacza to, że model DB2 został usunięty lub ponownie zbudowany po zbudowaniu modelu IBM SPSS Modeler. Więcej informacji można znaleźć w temacie “Karta serwera modelu użytkowego ISW” na stronie 70.

Lista modeli bazy danych

W produkcie IBM SPSS Modeler dostępne jest okno dialogowe, w którym można wyświetlić listę modeli zapisanych w produkcie IBM InfoSphere Warehouse Data Mining. W tym oknie modele można również usuwać. Dostęp do tego

okna można uzyskać z okna dialogowego IBM Helper Applications, a także z okien dialogowych do budowania, przeglądania i okna dialogowego Zastosuj dla węzłów dotyczących produktu IBM InfoSphere Warehouse Data Mining. Dla każdego modelu wyświetlane są następujące informacje:

- Nazwa modelu (używana w celu sortowania listy).
- Informacje o modelu (informacje o kluczu modelu pochodzące z klucza losowego, który został wygenerowany, gdy produkt IBM SPSS Modeler zbudował model).
- Typ modelu (tabela DB2, w której produkt IBM InfoSphere Warehouse Data Mining zapisał model).

Przeglądanie modeli

Narzędzie Visualizer jest jedyną metodą przeglądania modeli InfoSphere Warehouse Data Mining. Narzędzie może być opcjonalnie instalowane wraz z produktem InfoSphere Warehouse Data Mining. Więcej informacji można znaleźć w temacie „Włączanie integracji z produktem IBM InfoSphere Warehouse” na stronie 50.

- Kliknij opcję **Widok**, aby uruchomić narzędzie Visualizer. Treści, jakie zostaną wyświetlone w narzędziu, są zależne od typu wygenerowanego węzła. Na przykład narzędzie Visualizer zwróci widok klas przewidywanych, jeśli zostanie uruchomione z modelu użytkowego ISW Decision Tree.
- Kliknij opcję **Wyniki testu** (tylko drzewa decyzyjne i sekwencja), aby uruchomić narzędzie Visualizer i wyświetlić ogólną jakość wygenerowanego modelu.

Eksportowanie modeli i generowanie węzłów

W przypadku modeli IBM InfoSphere Warehouse Data Mining można wykonywać operacje importu i eksportu kodu PMML. Eksportowany kod PMML jest oryginalnym kodem PMML wygenerowanym przez produkt IBM InfoSphere Warehouse Data Mining. Funkcja eksportu zwraca model w formacie PMML.

Podsumowanie i strukturę modelu można wyeksportować do plików tekstowych i plików HTML. W razie potrzeby można generować odpowiednie węzły filtrowania, selekcji i wycień. Więcej informacji zawiera sekcja „Eksportowanie modeli” w dokumentacji *IBM SPSS Modeler — podręcznik użytkownika*.

Ustawienia węzłów wspólne dla wszystkich algorytmów

Następujące ustawienia są wspólne dla wielu algorytmów dostępnych w produkcie IBM InfoSphere Warehouse Data Mining:

Zmienne przewidywane i predyktory. Zmienną przewidywaną i predyktory można określić, korzystając z węzła typu lub ręcznie na karcie Zmienne w węźle konstruktora modeli, czyli w sposób standardowy dla produktu IBM SPSS Modeler.

Źródło danych ODBC. To ustawienie umożliwia użytkownikowi przesłonięcie domyślnego źródła danych ODBC dla bieżącego modelu. (Domyślne jest określone w oknie dialogowym Aplikacje pomocnicze. Więcej informacji można znaleźć w temacie „Włączanie integracji z produktem IBM InfoSphere Warehouse” na stronie 50.)

Opcje na karcie Serwer w ISW

Istnieje możliwość określenia połączenia DB2 używanego do przesyłania danych na potrzeby modelowania. W razie potrzeby można wybrać połączenie na karcie Serwer dla każdego węzła modelowania, aby przesłonić domyślne połączenie z DB2 określone w oknie dialogowym Aplikacje pomocnicze. Więcej informacji można znaleźć w temacie „Włączanie integracji z produktem IBM InfoSphere Warehouse” na stronie 50.

Połączenie używane na potrzeby modelowania może być takie samo, jak połączenie używane w węźle źródłowym dla strumienia. Na przykład może istnieć strumień, który uzyskuje dostęp do danych z jednej z baz danych DB2, pobiera dane do produktu IBM SPSS Modeler w celu kasowania lub w celu przeprowadzenia innych manipulacji, a następnie przesyła te dane do innej bazy danych DB2 na potrzeby modelowania.

Nazwa źródła danych ODBC jest osadzana w każdym strumieniu IBM SPSS Modeler. Jeśli strumień utworzony w jednym hoście zostanie wykonany na innym hoście, wówczas nazwa źródła danych musi być taka sama na każdym hoście. Alternatywnie inne źródło danych można wybrać na karcie Serwer w każdym węźle źródłowym lub węźle modelowania.

Następujące opcje umożliwiają uzyskiwanie informacji zwrotnych podczas budowania modelu:

- **Włącz sprzężenie zwrotne.** Wybierz tę opcję, aby uzyskiwać informacje zwrotne podczas budowania modelu (domyślnie ta funkcja jest wyłączona).
- **Przedział sprzężenia zwrotnego (w sekundach).** Określ częstotliwość, z jaką produkt IBM SPSS Modeler będzie otrzymywać informacje zwrotne na temat postępu budowania modelu.

Aktywuj opcje InfoSphere Warehouse Data Mining. Wybierz tę opcję, aby aktywować przycisk **Opcje zaawansowane**, który umożliwi określenie szeregu opcji zaawansowanych, np. limitu pamięci i niestandardowego kodu SQL. Więcej informacji można znaleźć w temacie “Opcje zaawansowane”.

Na karcie Serwer dla wygenerowanego węzła dostępna jest opcja przeznaczona do sprawdzania spójności poprzez zapisanie identycznego wygenerowanego klucza tekstowego w modelu IBM SPSS Modeler oraz w modelu DB2. Więcej informacji można znaleźć w temacie “Karta serwera modelu użytkowego ISW” na stronie 70.

Opcje zaawansowane

Karta Serwer dla wszystkich algorytmów zawiera pole wyboru, które umożliwia aktywację opcji zaawansowanych modelowania w ISW. Kliknięcie przycisku **Opcje zaawansowane** powoduje wyświetlenie okna dialogowego Opcje zaawansowane ISW, w którym dostępne są opcje:

- Limit pamięci.
- Inne opcje zaawansowane.
- Zgłębianie danych z użyciem SQL użytkownika.
- Dane logiczne SQL użytkownika.
- Ustawienia zgłębiania SQL.

Limit pamięci. Ogranicza zużycie pamięci przez algorytm budowania modelu. Należy zwrócić uwagę na to, że standardowo opcja zaawansowana ustawia limit dotyczący liczby wartości dyskretnych w danych jakościowych.

Inne opcje zaawansowane. Umożliwia określanie dowolnych opcji zaawansowanych w postaci wiersza komend. Dotyczy to tylko konkretnych modeli oraz rozwiązań. Sposób korzystania z poszczególnych opcji może być różny i jest uzależniony od implementacji lub rozwiązania. W celu zdefiniowania zadania budowania modelu można ręcznie rozszerzyć kod SQL wygenerowany przez produkt IBM SPSS Modeler.

Zgłębianie danych z użyciem SQL użytkownika. Istnieje możliwość dodawania wywołań metod w celu modyfikowania obiektu `DM_MiningData`. Na przykład wprowadzenie poniższego kodu SQL powoduje dodanie filtru na podstawie zmiennej o nazwie *Partycja* do danych używanych w ramach budowania modelu:

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

Dane logiczne SQL użytkownika. Istnieje możliwość dodawania wywołań metod w celu modyfikowania obiektu `DM_LogicalDataSpec`. Na przykład poniższy kod SQL usuwa zmienną z zestawu zmiennych używanych do budowania modelu:

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

Ustawienia zgłębiania SQL. Istnieje możliwość dodawania wywołań metod w celu modyfikowania obiektów `DM_ClasSettings/DM_RuleSettings/DM_ClusSettings/DM_RegSettings`. Na przykład poniższy kod SQL oznacza dla produktu IBM InfoSphere Warehouse Data Mining instrukcję ustawienia zmiennej *Partycja* w stan aktywny (co oznacza, że jest zawsze uwzględniana w modelu wynikowym):

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


Opcje kosztów ISW

Na karcie Koszty można dostosować koszty błędnej klasyfikacji, co pozwala na określenie względnej ważności różnych rodzajów błędów predykcji.

W niektórych kontekstach pewne błędy są bardziej kosztowne od innych. Przykładowo bardziej kosztowne może być sklasyfikowanie osób składających wnioski kredytowe z wysokim poziomem ryzyka jako osób z niskim poziomem ryzyka (jeden rodzaj błędu) niż sklasyfikowanie osób składających wnioski z niskim poziomem ryzyka jako osób z wysokim poziomem ryzyka (inny rodzaj błędu). Koszty błędnej klasyfikacji umożliwiają określenie względnej ważności różnych rodzajów błędów predykcji.

Kosztami błędnej klasyfikacji zwykle są wagi zastosowane do określonych danych wynikowych. Wagi te są uwzględniane w modelu i rzeczywiście mogą zmienić predykcję (jako sposób ochrony przed kosztownymi błędami).

Z wyjątkiem modeli C5.0 koszty błędnej klasyfikacji nie mają zastosowania podczas oceniania modelu i nie są brane pod uwagę podczas rangowania lub porównywania modeli za pomocą węzła Auto Klasyfikacja, wykresu ewaluacyjnego lub węzła analizy. Model, który uwzględnia koszty, może nie wygenerować mniejszej liczby błędów niż ten, który ich nie uwzględnia, i może nie mieć wyższej rangi pod względem całkowitej dokładności, ale prawdopodobnie lepiej sprawdzi się w warunkach praktycznych, ponieważ generuje błędy *mniej kosztowne*.

Macierz kosztów przedstawia koszty dla każdej możliwej kombinacji przewidywanej kategorii rzeczywistej. Domyślnie wszystkie koszty błędnej klasyfikacji są ustawione na wartość 1,0. Aby wprowadzić niestandardowe wartości kosztów, należy wybrać opcję **Stosuj koszty błędnej klasyfikacji** i wprowadzić do macierzy kosztów niestandardowe wartości.

Aby zmienić koszt błędnej klasyfikacji, należy zaznaczyć komórkę odpowiadającą odpowiedniej kombinacji wartości przewidywanych i rzeczywistych, usunąć istniejącą zawartość komórki i wprowadzić do niej żądany koszt. Koszty nie są automatycznie symetryczne. Przykładowo, jeśli koszt błędnej klasyfikacji *A* jako *B* zostanie ustawiony na 2,0, to koszt błędnej klasyfikacji *B* jako *A* nadal będzie miał domyślną wartość 1,0, chyba że zostanie ona również jawnie zmieniona.

ISW Decision Tree

Modele drzew decyzyjnych pozwalają tworzyć systemy klasyfikacyjne, które dokonują predykcji lub grupowania przyszłych obserwacji na podstawie zestawu reguł decyzyjnych. Mając dane podzielone na interesujące nas klasy (np. kredyty o wysokim ryzyku kontra kredyty o niskim ryzyku, abonenci kontra użytkownicy prepaid, głosujący kontra niegłosujący, typy bakterii), możemy wykorzystać te dane do budowania reguł klasyfikujących stare lub nowe obserwacje z maksymalną dokładnością. Na przykład możemy zbudować drzewo klasyfikujące ryzyko kredytowe lub zamiar zakupu na podstawie wieku i innych czynników.

Algorytm ISW Decision Tree buduje drzewa klasyfikacji w jakościowych danych wejściowych. Wynikowe drzewo decyzyjne jest binarne. W celu zbudowania modelu można zastosować różne ustawienia, w tym dotyczące kosztów błędnej klasyfikacji.

Narzędzie ISW Visualizer jest jedyną metodą przeglądania modeli InfoSphere Warehouse Data Mining.

Opcje modelu ISW Decision Tree

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu, w przypadkach gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Użyj danych podzielonych na podzbiory. Jeśli definiujesz zmienną dzielącą na podzbiory, wybierz opcję **Użyj danych podzielonych na podzbiory**.

Wykonaj uruchomienie testowe. Możesz zdecydować, aby wykonać uruchomienie testowe. Po wybraniu tej opcji uruchomienie testowe produktu IBM InfoSphere Warehouse Data Mining jest wykonywane po zbudowaniu modelu w podzbiorze uczącym. To powoduje przejście względem podzbioru testowego w celu ustalenia jakości modelu, wykresów przyrostów itp.

Maksymalna głębokość drzewa. Możliwe jest określenie maksymalnej głębokości drzewa. Ta opcja ogranicza głębokość drzewa do określonej liczby poziomów. Jeśli ta opcja nie zostanie wybrana, żaden limit nie będzie obowiązywał. Aby uniknąć powstawania nadmiernie złożonych modeli, wartość wyższa niż 5 rzadko jest polecana.

Opcje zaawansowane ISW Decision Tree

Maksymalna czystość. Ta opcja ustawia maksymalną czystość dla węzłów wewnętrznych. Jeśli podział węzła powoduje, że jeden element podrzędny przekracza określoną miarę czystości (na przykład więcej niż 90% obserwacji należy do określonej kategorii), wówczas taki węzeł nie jest dzielony.

Minimum obserwacji na węzeł wewnętrzny. Jeśli podział węzła powoduje powstanie węzła z mniejszą liczbą obserwacji niż określone minimum, węzeł nie zostanie podzielony.

ISW Association

Węzła ISW Association można używać, aby znajdować reguły asocjacyjne wśród elementów istniejących w zestawie grup. Reguły asocjacyjne kojarzą konkretny wniosek (np. decyzję o zakupie konkretnego produktu) ze zbiorem warunków (na przykład z zakupem kilku innych produktów).

Reguły asocjacyjne można uwzględnić w modelu lub wykluczyć z modelu, określając odpowiednie **ograniczenia**. Jeśli konkretna zmienna wejściowa zostanie uwzględniona, wówczas w modelu zostaną uwzględnione reguły asocjacyjne zawierające co najmniej jeden z określonych elementów. Jeśli zmienna wejściowa zostanie wykluczona, wówczas reguły asocjacyjne zawierające dowolne z określonych elementów zostaną usunięte z wyników.

Algorytmy ISW Association i ISW Sequence mogą korzystać z **systematyki**. Systematyki odwzorowują pojedyncze wartości na koncepcje wyższego poziomu. Na przykład długopisy i pióra mogą być mapowane na kategorię papeteria.

Każda reguła asocjacyjna ma pojedynczego następnika (wniosek) i wielu poprzedników (zestaw warunków). Przykład:

[Chleb, Dżem] △ [Masło]

[Chleb, Dżem]
△ [Margaryna]

W tym przypadku Chleb i Dżem są poprzednikami (co jest również określane jako **treść reguły**), a Masło lub Margaryna są przykładami następników (co jest również określane jako **nagłówek reguły**). Pierwsza reguła wskazuje, że osoba, która kupiła chleb i dżem, jednocześnie kupiła masło. Druga reguła identyfikuje klienta, który, kupując tę samą kombinację (chleb i dżem), kupił także margarynę w trakcie jednej wizyty w sklepie.

Narzędzie Visualizer jest jedyną metodą przeglądania modeli InfoSphere Warehouse Data Mining.

Opcje zmiennej ISW Association

Na karcie Zmienne można wskazać zmienne, które mają zostać użyte do budowy modelu.

Aby możliwe było zbudowanie modelu, konieczne jest określenie, które zmienne mają być używane jako zmienne przewidywane, a które jako dane wejściowe. We wszystkich węzłach modelowania (z kilkoma wyjątkami) stosowane są informacje na temat zmiennych z wcześniejszego węzła Typ. Ustawieniem domyślnym korzystania z węzła Typ jest wybieranie zmiennych wejściowych i przewidywanych, a jedynym innym ustawieniem, które można zmienić na tej karcie, jest układ tabeli w przypadku danych nietransakcyjnych.

Użyj ustawień węzła Typ. Ta opcja określa sposób użycia informacji o zmiennej z poprzedzającego węzła Typ. Jest to ustawienie domyślne.

Użyj ustawień niestandardowych. Ta opcja określa sposób użycia informacji o zmiennej wprowadzonej w tym miejscu zamiast podanej w dowolnym poprzedzającym węźle Typ. Po wybraniu tej opcji określ poniższe zmienne odpowiednio do potrzeb.

Użyj formatu transakcyjnego. To pole wyboru należy zaznaczyć, jeśli dane źródłowe są dostępne w **formacie transakcyjnym**. Rekordy w tym formacie mają dwie zmienne — jedną dla identyfikatora, a drugą dla treści. Każdy rekord reprezentuje pojedynczą transakcję lub element, a skojarzone elementy są powiązane, ponieważ mają ten sam identyfikator. Usuń zaznaczenie tego pola wyboru, jeśli dane mają **format tabel**, w których elementy są reprezentowane przez osobne flagi, a każda zmienna flagi reprezentuje obecność lub brak konkretnego elementu, a każdy rekord reprezentuje pełny zestaw skojarzonych elementów.

- **Identyfikator.** W przypadku danych transakcyjnych należy wybrać z listy zmienną identyfikacyjną. Jako zmienna identyfikacyjna mogą być używane zmienne numeryczne lub symboliczne. Każda unikalna wartość tej zmiennej powinna wskazywać na określoną jednostkę analizy. Na przykład w aplikacji do obsługi koszyka zakupów każdy identyfikator może reprezentować jednego klienta. W przypadku aplikacji do analizy dzienników sieciowych każdy identyfikator może reprezentować komputer (wg adresu IP) lub użytkownika (wg danych logowania).
- **Zawartość.** Należy określić zmienną lub zmienne zawartości dla modelu. Zmienne te zawierają interesujące elementy w procesie modelowania sekwencji. Gdy dane są dostępne w formacie transakcyjnym, można określić pojedynczą zmienną nominalną.

Użyj formatu tabelarycznego. Jeśli dane źródłowe są dostępne w formacie tabelarycznym, usuń zaznaczenie pola wyboru **Użyj formatu transakcyjnego**.

- **Zmienne wejściowe.** Umożliwia wybór zmiennej wejściowej lub kilku zmiennych. Działanie jest podobne jak w przypadku ustawienia roli zmiennej na *Zmienna wejściowa* w węźle Typ.
- **Podział.** To pole umożliwi określenie zmiennej używanej do podziału danych na osobne próby do uczenia, testowania i walidacji podczas budowania modelu. Korzystając z jednej próby do generowania modelu oraz innej do testowania go, można uzyskać wiarygodne informacje o tym, jak dobrze model pozwala uogólnić większe zbiory danych, podobne do danych bieżących. Jeśli korzystając z węzłów Typ lub Partycja, zdefiniowano wiele zmiennych dzielących na podzbiory, na karcie Zmienne każdego węzła modelowania korzystającego z tego podziału na podzbiory należy wybrać jedną zmienną dzielącą na podzbiory. (Jeśli obecna jest tylko jedna zmienna dzieląca na podzbiory, jest ona automatycznie używana po aktywowaniu dzielenia). Należy również pamiętać, że aby zastosować wybrany podział w analizie, dzielenie musi być również włączone na karcie Opcje modelu danego węzła. (Usunięcie zaznaczenia tej opcji umożliwia wyłączenie podziału bez zmiany ustawień zmiennych).

Wygląd tabeli danych nietransakcyjnych. W przypadku danych tabelarycznych można wybrać standardowy układ tabeli (domyślny) lub układ z ograniczoną długością elementów.

W układzie domyślnym liczba kolumn jest określona przez łączną liczbę skojarzonych elementów.

Tabela 6. Domyślny układ tabeli.

ID grupy	Rachunek bieżący	Rachunek oszczędnościowy	Karta kredytowa	Pożyczka	Rachunek powierniczy
Smith	Y	Y	Y	-	-
Jackson	Y	-	Y	Y	Y
Douglas	Y	-	-	-	Y

W układzie z ograniczoną długością elementów liczba kolumn jest określona przez najwyższą liczbę skojarzonych elementów w dowolnych wierszach.

Tabela 7. Układ tabeli z ograniczoną długością elementów.

ID grupy	Element1	Element2	Element3	Element4
Smith	rachunek bieżący	rachunek oszczędnościowy	karta kredytowa	-

Tabela 7. Układ tabeli z ograniczoną długością elementów (kontynuacja).

ID grupy	Element1	Element2	Element3	Element4
Jackson	rachunek bieżący	karta kredytowa	pożyczka	rachunek powierniczy
Douglas	rachunek bieżący	rachunek powierniczy	-	-

Opcje modelu asocjacyjnego ISW

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu, w przypadkach gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Użyj danych podzielonych na podzbiory. Jeśli zdefiniowano zmienną dzielącą na podzbiory, ta opcja umożliwia użycie podczas budowania modelu wyłącznie danych z podzbioru uczącego.

Minimalne pokrycie reguł (%). Minimalny poziom obsługi dla reguł asocjacji lub reguł sekwencyjnych. W modelu uwzględniane są tylko reguły, które osiągają co najmniej ten poziom obsługi. Wartość jest obliczana z równania $A/B*100$, gdzie A jest liczbą grup zawierających wszystkie elementy występujące w regule, a B jest łączną liczbą wszystkich uwzględnianych grup. W celu skoncentrowania się na bardziej typowych asocjacjach lub sekwencjach należy zwiększyć wartość tego ustawienia.

Minimalna ufność reguły (%). Minimalny poziom ufności dla reguł asocjacji lub reguł sekwencyjnych. W modelu uwzględniane są tylko reguły, które osiągają co najmniej ten poziom ufności. Wartość jest obliczana z równania $m/n*100$, gdzie m jest liczbą grup zawierających połączony nagłówek reguły (następnik) i treść reguły (poprzednik), a n jest liczbą grup zawierających treść reguły. W przypadku uzyskiwania zbyt wielu asocjacji lub sekwencji albo sekwencji mało interesujących należy spróbować zwiększyć wartość tego ustawienia. W przypadku uzyskiwania zbyt niewielu asocjacji lub sekwencji należy spróbować obniżyć wartość tego ustawienia.

Maksymalny rozmiar reguły. Maksymalna liczba elementów dozwolonych w regule z uwzględnieniem elementu następnika. Jeśli asocjacje lub sekwencje są relatywnie krótkie, można zmniejszyć wartość tego ustawienia w celu przyspieszenia budowy zestawu.

Uwaga: oceniane są tylko węzły, w których dane wejściowe mają format transakcyjny; formaty tablicy prawdy (dane tabelaryczne) pozostają surowe.

Opcje zaawansowane ISW Association

Na karcie Zaawansowane w węzle asocjacji można określić reguły asocjacyjne, które zostaną uwzględnione w wynikach albo wykluczone z wyników. Jeśli podejmiesz decyzję o uwzględnieniu konkretnych elementów, wówczas reguły zawierające co najmniej jeden z określonych elementów zostaną uwzględnione w modelu. Jeśli podejmiesz decyzję o wykluczeniu konkretnych elementów, wówczas reguły zawierające dowolne z tych elementów zostaną usunięte z wyników.

Gdy wybrana jest opcja **Stosuj ograniczenia elementów**, wszelkie elementy dodane do listy ograniczeń zostaną uwzględnione w wynikach lub wykluczone z wyników, jest to uzależnione od ustawienia **Typ ograniczenia**

Typ ograniczenia. Wybierz dołączanie lub wykluczanie z wyników tych reguł asocjacyjnych, które zawierają określone elementy.

Edytuj ograniczenia. Aby dodać element do listy ograniczonych elementów, wybierz element na liście elementów i kliknij przycisk ze strzałką w prawo.

Opcje systematyki ISW

Algorytmy ISW Association i ISW Sequence mogą korzystać z **systematyki**. Systematyki odwzorowują pojedyncze wartości na koncepcje wyższego poziomu. Na przykład długopisy i pióra mogą być mapowane na kategorię papeteria.

Na karcie Systematyka można zdefiniować mapy kategorii, aby wyrazić systematykę w danych. Na przykład systematyka może tworzyć dwie kategorie (Towary podstawowe i Towary luksusowe), a następnie przypisywać podstawowe elementy do każdej z tych kategorii. Na przykład wino jest przypisane do kategorii Towary luksusowe, a chleb jest przypisany do kategorii Towary podstawowe. Ta systematyka ma strukturę nadrzędny-podrzędny, którą przedstawia poniższa tabela.

Tabela 8. Przykład struktury systematyki

Elementy podrzędne	Elementy nadrzędne
wine	Towary luksusowe
bread	Towary podstawowe

Gdy systematyka obowiązuje, można zbudować model asocjacyjny lub model sekwencji zawierający reguły, które obejmują kategorie, a także elementy podstawowe.

Uwaga: w celu aktywacji opcji na tej karcie dane źródłowe muszą być w formacie transakcyjnym i należy wybrać opcję **Użyj formatu transakcyjnego** na karcie **Zmienne**, a następnie wybrać opcję **Używaj systematyki** na tej karcie.

Nazwa tabeli. Ta opcja określa nazwę tabeli DB2, w której zapisane zostaną szczegóły systematyki.

Kolumna elementów podrzędnych. Ta opcja określa nazwę kolumny elementów podrzędnych w tabeli systematyki. Kolumna elementów podrzędnych zawiera nazwy elementów lub nazwy kategorii.

Kolumna elementów nadrzędnych. Ta opcja określa nazwę kolumny elementów nadrzędnych w tabeli systematyki. Kolumna elementów nadrzędnych zawiera nazwy kategorii.

Załaduj szczegóły do tabeli. Tę opcję należy wybrać, jeśli informacje dotyczące systematyki w produkcie IBM SPSS Modeler powinny zostać wprowadzone do tabeli systematyki podczas budowania modelu. Należy zwrócić uwagę na to, że jeśli tabela systematyki już istnieje, wówczas jest usuwana. Informacje o systematyce są zapisywane w węzle budowania modelu i są edytowane przy użyciu przycisków **Edytuj kategorie** i **Edytuj systematykę**.

Edytor kategorii

Okno dialogowe edytora kategorii umożliwia dodawanie kategorii i usuwanie ich z sortowanej listy.

W celu dodania kategorii wpisz jej nazwę do pola **Nowa kategoria** i kliknij przycisk strzałki, aby przesunąć ją do listy **Kategorie**.

W celu usunięcia kategorii wybierz ją na liście **Kategorie** i kliknij sąsiadujący przycisk **Usuń**.

Edytor systematyki

W oknie edytora systematyki można ustawiać podstawowe elementy zdefiniowane w danych, a także zestaw kategorii do połączenia w celu zbudowania systematyki. W celu dodawania pozycji do systematyki wybierz co najmniej jeden element lub co najmniej jedną kategorię z listy po lewej stronie i co najmniej jedną kategorię z listy po prawej stronie, a następnie kliknij przycisk ze strzałką. Zwróć uwagę na to, że jeśli dodanie jakichkolwiek pozycji do systematyki spowoduje konflikt (na przykład określenie **cat1 -> cat2** i przeciwności **cat2 -> cat1**), wówczas te pozycje nie zostaną dodane.

ISW Sequence

Węzeł Sekwencje wykrywa wzorce w sekwencyjnych lub zorientowanych czasowo danych, w następującej postaci **bread -> cheese**. Elementy w kolejności to **zbiory elementów**, które tworzą pojedynczą transakcję. Na przykład, jeśli dana osoba idzie do sklepu i kupuje chleb i mleko, a następnie po kilku dniach wraca do sklepu i kupuje ser, aktywność zakupową tej osoby mogą reprezentować dwa zbiory elementów. Pierwszy zbiór elementów zawiera chleb i mleko, a drugi zawiera ser. **Sekwencja** to lista zbiorów elementów z tendencją do występowania w przewidywalnej kolejności. Węzeł Sekwencje wykrywa powtarzające się sekwencje i tworzy wygenerowany węzeł modelu, którego można użyć do predykcji.

Funkcję eksploracji reguł sekwencyjnych można wykorzystać w różnych strefach działalności biznesowej. Na przykład w branży handlu detalicznego można znajdować typowe serie zakupów. Te serie przedstawiają różne kombinacje klientów, produktów i czasu zakupu. Takie informacje umożliwiają zidentyfikowanie potencjalnych klientów — odbiorców konkretnego produktu, którzy jeszcze takiego produktu nie kupili. Ponadto można oferować produkty potencjalnym klientom w odpowiednim czasie.

Sekwencja jest uporządkowanym zestawem zbiorów elementów. Sekwencje zawierają następujące poziomy grupowania:

- Zdarzenia, do których dochodzi jednocześnie, tworzą pojedynczą transakcję lub zbiór elementów.
- Każdy element lub zbiór elementów należy do grupy transakcji. Na przykład zakupiony artykuł należy do klienta, konkretne kliknięcie na stronie należy do osoby surfującej po Internecie, a podzespół należy do produkowanego samochodu. Kilka zbiorów elementów, do których dochodzi w różnym czasie i które należą do tej samej grupy transakcji, tworzą sekwencję.

Opcje modelu ISW Sequence

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu, w przypadkach gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Użyj danych podzielonych na podzbiory. Jeśli zdefiniowano zmienną dzielącą na podzbiory, ta opcja umożliwia użycie podczas budowania modelu wyłącznie danych z podzbioru uczącego.

Minimalne pokrycie reguł (%). Minimalny poziom obsługi dla reguł asocjacji lub reguł sekwencyjnych. W modelu uwzględniane są tylko reguły, które osiągają co najmniej ten poziom obsługi. Wartość jest obliczana z równania $A/B*100$, gdzie A jest liczbą grup zawierających wszystkie elementy występujące w regule, a B jest łączną liczbą wszystkich uwzględnianych grup. W celu skoncentrowania się na bardziej typowych asocjacjach lub sekwencjach należy zwiększyć wartość tego ustawienia.

Minimalna ufność reguły (%). Minimalny poziom ufności dla reguł asocjacji lub reguł sekwencyjnych. W modelu uwzględniane są tylko reguły, które osiągają co najmniej ten poziom ufności. Wartość jest obliczana z równania $m/n*100$, gdzie m jest liczbą grup zawierających połączony nagłówek reguły (następnik) i treść reguły (poprzednik), a n jest liczbą grup zawierających treść reguły. W przypadku uzyskiwania zbyt wielu asocjacji lub sekwencji albo sekwencji mało interesujących należy spróbować zwiększyć wartość tego ustawienia. W przypadku uzyskiwania zbyt niewielu asocjacji lub sekwencji należy spróbować obniżyć wartość tego ustawienia.

Maksymalny rozmiar reguły. Maksymalna liczba elementów dozwolonych w regule z uwzględnieniem elementu następnika. Jeśli asocjacje lub sekwencje są relatywnie krótkie, można zmniejszyć wartość tego ustawienia w celu przyspieszenia budowy zestawu.

Uwaga: oceniane są tylko węzły, w których dane wejściowe mają format transakcyjny; formaty tablicy prawdy (dane tabelaryczne) pozostają surowe.

Opcje zaawansowane ISW Sequence

Możesz określić reguły sekwencyjne, które zostaną uwzględnione w wynikach albo wykluczone z wyników. Jeśli podejmiesz decyzję o uwzględnieniu konkretnych elementów, wówczas reguły zawierające co najmniej jeden z określonych elementów zostaną uwzględnione w modelu. Jeśli podejmiesz decyzję o wykluczeniu konkretnych elementów, wówczas reguły zawierające dowolne z tych elementów zostaną usunięte z wyników.

Gdy wybrana jest opcja **Stosuj ograniczenia elementów**, wszelkie elementy dodane do listy ograniczeń zostaną uwzględnione w wynikach lub wykluczone z wyników, jest to uzależnione od ustawienia **Typ ograniczenia**

Typ ograniczenia. Wybierz dołączanie lub wykluczanie z wyników tych reguł asocjacyjnych, które zawierają określone elementy.

Edytuj ograniczenia. Aby dodać element do listy ograniczonych elementów, wybierz element na liście elementów i kliknij przycisk ze strzałką w prawo.

ISW Regression

Węzeł ISW Regression obsługuje następujące algorytmy regresji:

- Transformacja (domyślny)
- Liniowy
- Wielomianowy
- RBF

Regresja z transformacją

Algorytm regresji z transformacją ISW buduje modele będące drzewami decyzyjnymi z równaniami regresji na liściach drzew. Należy zwrócić uwagę, że w narzędziu IBM Visualizer struktura tych modeli nie będzie wyświetlana.

Przeglądarka produktu IBM SPSS Modeler przedstawia ustawienia i adnotacje. Jednak przeglądanie struktury modelu jest niemożliwe. W strukturze jest niewiele ustawień budowania, które może konfigurować użytkownik.

Regresja liniowa

Algorytm regresji liniowej ISW przyjmuje założenie, że między zmiennymi wyjaśniającymi a zmienną przewidywaną istnieje zależność liniowa. Generuje modele, które reprezentują równania. Zgodnie z oczekiwaniami wartość przewidywana powinna się różnić od wartości obserwowanej, ponieważ równanie regresji jest przybliżeniem zmiennej przewidywanej. Ta różnica jest nazywana resztą.

Modelowanie w produkcie IBM InfoSphere Warehouse Data Mining rozpoznaje zmienne, dla których nie istnieje wartość wyjaśniająca. W celu ustalenia, czy dla zmiennej istnieje wartość wyjaśniająca, algorytm regresji liniowej przeprowadza testy statystyczne, a także autonomiczny wybór zmiennej. Jeśli znasz zmienne, dla których nie istnieje wartość wyjaśniająca, wówczas możesz automatycznie wybrać podzbiór zmiennych wyjaśniających, aby zapewnić krótsze czasy wykonania.

Algorytm regresji liniowej udostępnia następujące metody automatycznego wybierania podzbiorów zmiennych wyjaśniających:

Regresja krokowa. W przypadku regresji krokowej należy określić minimalny poziom istotności. Algorytm regresji liniowej wykorzysta tylko te zmienne, których poziom istotności jest wyższy niż podana wartość.

Regresja R-kwadrat. Metoda regresji R-kwadrat wskazuje model optymalny poprzez optymalizację miary jakości modelu. Używana jest jedna z poniższych miar jakości:

- Kwadrat współczynnika korelacji Pearsona.
- Skorygowany kwadrat współczynnika korelacji Pearsona.

Domyślnie algorytm regresji liniowej automatycznie wybiera podzbiory zmiennych wyjaśniających, używając skorygowanego kwadratu współczynnika korelacji Pearsona, aby zoptymalizować jakość modelu.

Regresja wielomianowa

Algorytm regresji wielomianowej ISW zakłada, że istnieje zależność wielomianowa. Model regresji wielomianowej jest równaniem, które składa się z następujących części:

- Maksymalny stopień regresji wielomianowej
- Przybliżenie zmiennej przewidywanej
- Zmienne wyjaśniające

Regresja RBF

Algorytm regresji RBF ISW przyjmuje założenie, że między zmiennymi wyjaśniającymi a zmienną przewidywaną istnieje zależność. Ta zależność może zostać wyrażona jako liniowa kombinacja funkcji Gaussa. Funkcje Gaussa to konkretne radialne funkcje bazowe (RBF).

Opcje modelu ISW Regression

Na karcie Model węzła ISW Regression można określić typ używanego algorytmu regresji, a także określić:

- Czy używane będą dane podzielone na podzbiory
- Czy wykonane zostanie uruchomienie testowe
- Limit dla wartości R^2
- Limit dla czasu wykonania

Użyj danych podzielonych na podzbiory. Jeśli zdefiniowano zmienną dzielącą na podzbiory, ta opcja umożliwia użycie podczas budowania modelu wyłącznie danych z podzbioru uczącego.

Metoda liczenia regresji. Wybierz typ regresji, jaki zamierzasz wykonać. Więcej informacji można znaleźć w temacie “ISW Regression” na stronie 63.

Wykonaj uruchomienie testowe. Możesz zdecydować, aby wykonać uruchomienie testowe. Po wybraniu tej opcji uruchomienie testowe produktu InfoSphere Warehouse Data Mining jest wykonywane po zbudowaniu modelu w podzbiorze uczącym. To powoduje przejście względem podzbioru testowego w celu ustalenia jakości modelu, wykresów przyrostów itp.

Ogranicz R kwadrat. Ta opcja określa maksymalny tolerowany błąd systematyczny (kwadrat współczynnika korelacji Pearsona, R^2). Ten współczynnik mierzy korelację między błędem predykcji w danych weryfikacyjnych i rzeczywistymi wartościami przewidywanymi. Ma wartość z zakresu od 0 (brak korelacji) do 1 (idealna korelacja dodatnia lub ujemna). Wartość zdefiniowana w tym miejscu ustawia górny limit dla akceptowalnego błędu systematycznego modelu.

Ogranicz czas wykonywania. Określ żądany maksymalny czas wykonania w minutach.

Opcje zaawansowane ISW Regression

Na karcie Zaawansowane węzła ISW Regression możesz określić liczbę opcji zaawansowanych dla regresji liniowej, wielomianowej lub regresji RBF.

Opcje zaawansowane dla regresji liniowej i wielomianowej

Ogranicz stopień wielomianu. Umożliwia ustawienie maksymalnego stopnia regresji wielomianowej. Jeśli maksymalny stopień regresji wielomianowej zostanie ustawiony na **1**, wówczas algorytm regresji wielomianowej będzie taki sam, jak algorytm regresji liniowej. Jeśli maksymalny stopień regresji wielomianowej zostanie ustawiony na wysoką wartość, wówczas algorytm regresji wielomianowej będzie wykazywał tendencje do przeuczenia. Oznacza to, że model wynikowy dokładnie przybliży dane uczące, jednak nie zapewnia przybliżenia, gdy jest stosowany względem danych, które nie są używane do uczenia.

Użyj wyrazu wolnego. Włączenie tej opcji wymusza przejście krzywej regresji przez początek układu współrzędnych. Oznacza to, że model nie będzie zawierał składnika stałego.

Użyj automatycznego wyboru predyktora. Gdy ta opcja jest włączona, algorytm podejmuje próbę ustalenia optymalnego podzbioru możliwych predyktorów, jeśli użytkownik nie określił minimalnego poziomu istotności.

Użyj minimalnego poziomu istotności. Gdy minimalny poziom istotności jest określony, wówczas używana jest regresja krokowa w celu określenia podzbioru możliwych predyktorów. Wkład w obliczenie modelu regresji wnoszą tylko zmienne niezależne, których istotność jest wyższa niż podana wartość.

Ustawienia zmiennych. W celu określenia opcji dla pojedynczych zmiennych wejściowych kliknij odpowiedni wiersz w kolumnie ustawień w tabeli Ustawienia zmiennych, a następnie wybierz opcję **<Określ ustawienia>**. Więcej informacji można znaleźć w temacie “Określanie ustawień zmiennych na potrzeby regresji”.

Opcje zaawansowane dla regresji RBF

Użyj wynikowej wielkości próby. Definiuje próbę 1-w-n na potrzeby weryfikacji i testowania modelu.

Użyj wejściowej wielkości próby. Definiuje próbę 1-w-n na potrzeby uczenia.

Użyj maksymalnej liczby środków. Maksymalna liczba ośrodków budowanych w każdym przebiegu. W trakcie jednego przebiegu liczba ośrodków może wzrosnąć dwukrotnie względem liczby początkowej, dlatego rzeczywista liczba ośrodków może być wyższa niż liczba określona.

Użyj minimalnego rozmiaru regionu. Minimalna liczba rekordów przypisanych do regionu.

Użyj maksymalnego przekazywania danych. Maksymalna liczba przebiegów przez dane wejściowe, jaką wykonuje algorytm. Jeśli ta wartość zostanie określona, musi być równa minimalnej liczby przebiegów lub większa od niej.

Użyj minimalnego przekazywania danych. Minimalna liczba przebiegów przez dane wejściowe, jaką wykonuje algorytm. Określ dużą wartość, pod warunkiem że ilość danych uczących jest wystarczająca i masz pewność, że istnieje dobry model.

Określanie ustawień zmiennych na potrzeby regresji

W oknie dialogowym Edytuj ustawienia regresji można określić zakres wartości dla pojedynczej zmiennej wejściowej dla regresji liniowej lub wielomianowej.

Wartość MIN. Minimalna wartość poprawna dla danej zmiennej wejściowej.

Wartość MAKS. Maksymalna wartość poprawna dla danej zmiennej wejściowej.

ISW Clustering

Funkcja eksploracji z grupowaniem wyszukuje dane wejściowe dla cech, które najczęściej występują powszechnie. Grupuje dane wejściowe w skupienia. Elementy należące do poszczególnych skupień mają podobne właściwości. Nie istnieją żadne wcześniejsze wyobrażenia tego, jakie wzorce istnieją w danych. Grupowanie jest procesem wykrywania.

Wzrost ISW Clustering umożliwia wybór następujących metod grupowania:

- Demograficzna
- Sieć Kohonena
- Rozszerzony BIRCH (ang. Balanced Iterative Reducing and Clustering using Hierarchies)

Algorytm **grupowania demograficznego** wykorzystuje technikę bazującą na rozkładzie. Grupowanie bazujące na rozkładzie zapewnia szybkie i naturalne grupowanie bardzo dużych baz danych. Liczba skupień jest wybierana automatycznie (możliwe jest określenie maksymalnej liczby skupień). Istnieje duża liczba parametrów, które może skonfigurować użytkownik.

Algorytm **grupowania Kohonena**, wykorzystuje technikę zależną od środka. Mapa elementów Kohonena podejmuje próbę umieszczenia środków skupień w miejscach, które minimalizują ogólną odległość między rekordami a środkami skupień. Oddzielność skupień nie jest uwzględniana. Wektory środkowe są ułożone w mapę o konkretnej liczbie kolumn i wierszy. Te wektory są połączone ze sobą w taki sposób, że korygowany jest nie tylko wektor wygrywający, który znajduje się najbliżej rekordu uczącego, ale także wektory w jego sąsiedztwie. Jednak im dalej znajdują się inne środki, tym mniej są korygowane.

Algorytm grupowania **rozszerzony BIRCH** wykorzystuje technikę zależną od rozkładu i podejmuje próbę minimalizowania ogólnej odległości między rekordami a ich skupieniami. Logarytm wiarygodności jako miara odległości jest stosowany domyślnie w celu określenia odległości między rekordem a skupieniem; alternatywnie można wybrać odległość euklidesową, jeśli wszystkie zmienne aktywne są liczbowe. Algorytm BIRCH wykonuje dwa niezależne kroki; najpierw układa rekordy wejściowe w drzewo składników grupowania w taki sposób, że podobne rekordy stają się częścią tych samych węzłów drzewa, a następnie grupuje liście drzewa w pamięci, aby wygenerować ostateczny wynik grupowania.

Opcje modelu ISW Clustering

Na karcie model węzła grupowania można określić metodę używaną do tworzenia skupień, a także inne powiązane opcje.

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu, w przypadkach gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Użyj danych podzielonych na podzbiory. Jeśli zdefiniowano zmienną dzielącą na podzbiory, ta opcja umożliwia użycie podczas budowania modelu wyłącznie danych z podzbioru uczącego.

Metoda aglomeracji. Wybierz metodę, której zamierzasz używać w celu tworzenia skupień: **Demograficzna**, **Kohonena** lub **Rozszerzony BIRCH**. Więcej informacji można znaleźć w temacie “ISW Clustering” na stronie 65.

Ogranicz liczbę grup. Ograniczenie liczby grup (skupień) powoduje oszczędności w postaci czasu, ponieważ zapobiega generowaniu zbyt małych skupień.

Liczba wierszy/Liczba kolumn. (Dotyczy tylko metody Kohonena) Określa liczbę wierszy i kolumn dla mapy elementów Kohonena. (Ta opcja jest dostępna tylko wówczas, gdy wybrana jest opcja **Ogranicz liczbę wykonań Kohonena**, a opcja **Ogranicz liczbę grup** nie jest zaznaczona).

Ogranicz liczbę wykonań Kohonena. (Dotyczy tylko metody Kohonena) Określa liczbę wykonań algorytmu grupowania względem danych podczas przebiegów uczących. Przy każdym wykonaniu wektory środkowe są regulowane, aby zminimalizować łączną odległość między środkami skupień a rekordami. Ponadto zmniejszany jest stopień regulacji wektorów. Podczas pierwszego wykonania regulacje są zgrubne. W ostatnim wykonaniu stopień regulacji środków jest raczej niewielki. Wykonywane są tylko niewielkie regulacje.

Miara odległości. (Dotyczy tylko metody Rozszerzony BIRCH) Należy wybrać miarę odległości od rekordu do skupienia, która będzie używana przez algorytm BIRCH. Można wybrać odległość, jaką jest logarytm wiarygodności, co jest opcją domyślną, albo odległość euklidesową. *Uwaga:* odległość euklidesową można wybrać tylko wówczas, gdy wszystkie zmienne aktywne są liczbowe.

Maksymalna liczba węzłów-liści. (Dotyczy tylko metody Rozszerzony BIRCH) Maksymalna liczba węzłów-liści, jaka ma istnieć w drzewie predyktorów między grupami. Drzewo predyktorów między grupami jest wynikiem pierwszego etapu działania algorytmu Rozszerzony BIRCH, w którym rekordy są układane w drzewie w taki sposób, że podobne rekordy należą do tego samego węzła-liścia. Czas działania algorytmu wzrasta, gdy zwiększa się liczba węzłów-liści. Domyślną wartością jest 1000.

Przejścia algorytmu BIRCH. (Dotyczy tylko metody Rozszerzony BIRCH) Określa liczbę wykonań algorytmu względem danych w celu sprecyzowania wyniku grupowania. Liczba wykonań wpływa na czas przetwarzania przebiegów uczących (ponieważ każde wykonanie wymaga pełnego skanowania danych), a także od jakości modelu. Niskie wartości powodują krótkie czasy przetwarzania; jednak mogą spowodować niższą jakość modeli. Wyższe wartości powodują dłuższy czas przetwarzania i zwykle zwracają lepsze modele. Przeciętnie 3 wykonania i więcej prowadzi do uzyskania dobrych wyników. Domyślna wartość to 3.

Opcje zaawansowane ISW Clustering

Na karcie Zaawansowane w węzle grupowania można określać zaawansowane opcje, takie jak progi podobieństwa, limity czasu wykonania oraz wagi zmiennych.

Ogranicz czas wykonywania. Zaznacz to pole wyboru, aby włączyć opcje, które umożliwiają kontrolowanie czasu tworzenia modelu. Czas można określić w minutach i/lub jako minimalny procent danych uczących przeznaczonych do przetworzenia. Ponadto w metodzie Birch można określić maksymalną liczbę węzłów-liści, które zostaną utworzone w drzewie CF.

Określ wartość graniczną podobieństwa. (Dotyczy tylko grupowania demograficznego) Dolny limit dla podobieństwa dwóch rekordów danych, które należą do tego samego skupienia. Na przykład wartość 0,25 oznacza, że rekordy z wartościami, które są w 25% podobne, prawdopodobnie zostaną przypisane do tego samego skupienia. Wartość 1,0 oznacza, że rekordy mogą się pojawić w tym samym skupieniu, pod warunkiem że są identyczne.

Ustawienia zmiennych. W celu określenia opcji dla pojedynczych zmiennych wejściowych kliknij odpowiedni wiersz w kolumnie ustawień w tabeli Ustawienia zmiennych, a następnie wybierz opcję **<Określ ustawienia>**.

Określanie ustawień zmiennych na potrzeby grupowania

W oknie dialogowym Edytuj ustawienia grupowania można określać opcje dotyczące poszczególnych zmiennych wejściowych.

Waga zmiennej. Przypisuje zmiennej większą lub mniejszą wagę podczas procesu budowania modelu. Na przykład, jeśli uważasz, że konkretna zmienna jest względnie mniej istotna dla modelu niż inne zmienne, zmniejsz wagę tej zmiennej w porównaniu z innymi zmiennymi.

Waga wartości. Przypisuje większą lub mniejszą wagę do konkretnych wartości danej zmiennej. Niektóre wartości zmiennych mogą być bardziej powszechne niż inne wartości. Zbieżność rzadkich wartości w zmiennej może być bardziej istotna dla skupienia niż zbieżność częstych wartości. W celu określania wag wartości dla tej zmiennej można posłużyć się jedną z poniższych metod (w każdym przypadku wartości radsze mają większą wagę, a wartości powszechne mają mniejszą wagę):

- **Logarytmiczny.** Przypisuje wagę do każdej wartości zgodnie z logarytmem jej prawdopodobieństwa w danych wejściowych.
- **Prawdopodobieństwa.** Przypisuje wagę do każdej wartości zgodnie z jej prawdopodobieństwem w danych wejściowych.

W przypadku każdej metody można również wybrać opcję **z kompensacją**, aby skompensować wagę wartości zastosowaną względem poszczególnych zmiennych. Jeśli skompensujesz wagi wartości, ogólna ważność zmiennej ważącej będzie równa ważności zmiennej nieważonej. Dzieje się tak bez względu na liczbę możliwych wartości. Wazenie skompensowane wpływa tylko na ważność względną zbieżności w zestawie możliwych wartości.

Stosuj skalę podobieństw. Zaznacz to pole wyboru, jeśli zamierzasz użyć skali podobieństw, aby kontrolować obliczenie miary podobieństwa dla zmiennej. Określ skalę podobieństw jako liczbę bezwzględną. Ta specyfikacja jest uwzględniana tylko w przypadku aktywnych zmiennych liczbowych. Jeśli skala podobieństw nie zostanie określona, wówczas używana będzie wartość domyślna (połowa odchylenia standardowego). W celu uzyskania większej liczby skupień należy zmniejszyć podobieństwo średnie między parami skupień o mniejsze skale podobieństw dla zmiennych liczbowych.

Traktowanie wartości odstających. Wartości odstające to wartości zmiennych, które znajdują się na zewnątrz zakresu wartości określonego dla zmiennej przez wartości **Wartość MIN** i **Wartość MAKS**. Możesz wybrać sposób postępowania z wartościami odstającymi w przypadku tej zmiennej.

- Ustawienie domyślne **brak** oznacza, że w przypadku wartości odstających nie są podejmowane żadne działania specjalne.
- Jeśli wybierzesz opcję **zamień wartością MIN lub MAKS**, wówczas wartość zmiennej mniejsza niż **Wartość MIN** lub większa niż **Wartość MAKS** zostanie zastąpiona odpowiednio wartościami MIN lub MAKS. W tym przypadku można ustawić wartości MIN i MAKS.

- Jeśli wybierzesz opcję **traktuj jako brak**, wówczas wartości odstające będą traktowane jako wartości brakujące i będą pomijane. W tym przypadku można ustawić wartości MIN i MAKS.

ISW Naive Bayes

Naive Bayes to dobrze znany algorytm stosowany w przypadku problemów z klasyfikacją. Model jest określany jako *naiwny*, ponieważ traktuje wszystkie proponowane zmienne predykcji jako niezależne od siebie. Naive Bayes to szybki, skalowalny algorytm, który oblicza prawdopodobieństwa warunkowe kombinacji atrybutów i atrybutu przewidywanego. Na podstawie danych uczących wyznaczane jest niezależne prawdopodobieństwo. To prawdopodobieństwo określa wiarygodność każdej klasy przewidywanej z uwzględnieniem wystąpienia każdej kategorii wartości z poszczególnych zmiennych wejściowych.

Algorytm klasyfikacji ISW Naive Bayes jest klasyfikatorem probabilistycznym. Jest oparty na modelach prawdopodobieństwa, w których obowiązują założenia silnej niezależności.

Opcje modelu ISW Naive Bayes

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu, w przypadkach gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Użyj danych podzielonych na podzbiory. Jeśli zdefiniowano zmienną dzielącą na podzbiory, ta opcja umożliwia użycie podczas budowania modelu wyłącznie danych z podzbioru uczącego.

Wykonaj uruchomienie testowe. Możesz zdecydować, aby wykonać uruchomienie testowe. Po wybraniu tej opcji uruchomienie testowe produktu IBM InfoSphere Warehouse Data Mining jest wykonywane po zbudowaniu modelu w podzbiorze uczącym. To powoduje przejście względem podzbioru testowego w celu ustalenia jakości modelu, wykresów przyrostów itp.

Wartość graniczna prawdopodobieństwa. Wartość graniczna prawdopodobieństwa definiuje prawdopodobieństwo dla dowolnych kombinacji predyktora i wartości przewidywanych, które nie znajdują się w danych uczących. Prawdopodobieństwo powinno mieścić się w zakresie od 0 do 1. Wartością domyślną jest 0,001.

ISW Logistic Regression

Regresja logistyczna, znana również pod nazwą regresji nominalnej, to technika statystyczna umożliwiająca klasyfikację rekordów na podstawie wartości zmiennych wejściowych. Jest ona analogiczna do regresji liniowej, ale algorytm ISW Logistic Regression przyjmuje flagową (binarną) zmienną przewidywaną, a nie liczbową.

Opcje modelu ISW Logistic Regression

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu, w przypadkach gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Użyj danych podzielonych na podzbiory. Jeśli zdefiniowano zmienną dzielącą na podzbiory, ta opcja umożliwia użycie podczas budowania modelu wyłącznie danych z podzbioru uczącego.

Wykonaj uruchomienie testowe. Możesz zdecydować, aby wykonać uruchomienie testowe. Po wybraniu tej opcji uruchomienie testowe produktu IBM InfoSphere Warehouse Data Mining jest wykonywane po zbudowaniu modelu w podzbiorze uczącym. To powoduje przejście względem podzbioru testowego w celu ustalenia jakości modelu, wykresów przyrostów itp.

ISW Time Series

Algorytm ISW Time Series umożliwia przewidywanie przyszłych wydarzeń na podstawie znanych wydarzeń z przeszłości.

Algorytmy szeregów czasowych (Time Series), podobnie jak typowe metody regresji, przewidują wartość liczbową. W odróżnieniu od typowych metod regresji predykcje szeregów czasowych dotyczą przyszłych wartości szeregów uporządkowanych. Te predykcje są często nazywane prognozami.

Algorytmy szeregów czasowych są algorytmami jednej zmiennej. Oznacza to, że zmienna niezależna jest kolumną czasu lub kolumną kolejności. Prognozy są oparte na przeszłych wartościach. Nie są oparte na innych kolumnach niezależnych.

Algorytmy szeregów czasowych różnią się od typowych algorytmów regresji, ponieważ nie tylko przewidują przyszłe wartości, ale również uwzględniają w prognozie cykle sezonowe.

Funkcja eksploracji Szereg czasowy zapewnia następujące algorytmy, które umożliwiają przewidywanie przyszłych trendów:

- Autoregresyjna zintegrowana średnia ruchoma (ARIMA)
- Wygładzanie wykładnicze
- Dekompozycja trendu sezonowego

Algorytm, który tworzy najlepszą prognozę na podstawie danych, jest zależny od różnych założeń dotyczących modelu. Wszystkie prognozy można wyliczyć jednocześnie. Algorytmy obliczają szczegółową prognozę, uwzględniając zachowania sezonowe oryginalnego szeregu czasowego. Jeśli zainstalowany jest klient IBM InfoSphere Warehouse Client, wówczas można użyć narzędzia Visualizer dotyczącego szeregów czasowych w celu oceny i porównania krzywych wyników.

Opcje zmiennych ISW Time Series

Czas. Wybierz zmienną wejściową, która zawiera szeregi czasowe. Musi to być zmienna o typie składowania Data, Godzina, Znacznik czasu, Liczba rzeczywista lub Liczba całkowita.

Użyj ustawień węzła Typ. Ta opcja stanowi dla węzła instrukcję o konieczności użycia informacji o zmiennej z wcześniejszego węzła Typ. Jest to ustawienie domyślne.

Użyj ustawień niestandardowych. Ta opcja stanowi dla węzła instrukcję o konieczności użycia informacji o zmiennej określonych w tym miejscu, a nie w żadnym wcześniejszym węźle Typ. Po wybraniu tej opcji określ poniższe zmienne odpowiednio do potrzeb.

Przewidywane. Wybierz co najmniej jedną zmienną przewidywaną. Działanie jest podobne jak w przypadku ustawienia roli zmiennej na wartość *Zmienna przewidywana* w węźle Typ.

Opcje modelu ISW Time Series

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu, w przypadkach gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Algorytmy prognozowania. Wybierz algorytmy, które będą używane na potrzeby modelowania. Można wybrać jedną z następujących opcji lub różne poniższe opcje:

- ARIMA
- Wygładzanie wykładnicze
- Dekompozycja trendu sezonowego

Czas końca prognozowania. Określ, czy czas końca prognozowania będzie obliczany automatycznie, czy będzie określany ręcznie.

Wartość pola czasu. Jeśli opcja **Czas końca prognozowania** jest ustawiona na wartość Ręczne, wówczas wprowadź czas końca prognozowania. Wartość, jaką można wprowadzić, jest zależna od typu zmiennej czasu; na przykład, jeśli

typ jest liczbą całkowitą, która reprezentuje godziny, wówczas można wprowadzić 48, aby zatrzymać prognozowanie po przetworzeniu danych z 48 godzin. Alternatywnie zmienna może zachęcać do wprowadzenia daty lub godziny jako wartości końca.

Opcje zaawansowane ISW Time Series

Użyj **wszystkich rekordów do budowania modelu**. Jest to ustawienie domyślne; wszystkie rekordy są analizowane po zbudowaniu modelu.

Do budowania modelu użyj podzbioru rekordów. Wybierz tę opcję, jeśli zamierzasz utworzyć model tylko z części dostępnych danych. Może to być konieczne na przykład, gdy istnieje bardzo duża ilość powtarzających się danych.

Wprowadź **Wartość czasu początkowego** oraz **Wartość czasu końcowego**, aby wskazać dane, które będą używane. Zwróć uwagę na to, że wartości, jakie można wprowadzać do tych pól, są zależne od typu zmiennej czasowej; na przykład może to być liczba godzin lub dni albo konkretne daty i godziny.

Metoda interpolacji dla brakujących wartości docelowych. Jeśli przetwarzasz dane z co najmniej jedną wartością brakującą, wybierz metodę obliczania takich wartości. Można wybrać jedną z następujących opcji:

- Liniowy
- Wykładnicze krzywe sklejące
- Sześciennicze krzywe sklejące

Wyświetlanie modeli ISW Time Series

Modele ISW Time Series są zwracane w postaci modelu surowego, który zawiera informacje wyodrębnione z danych, ale nie jest przeznaczony do bezpośredniego generowania predykcji.



Rysunek 3. Ikona modelu surowego

Jeśli zainstalowany jest klient IBM InfoSphere Warehouse Client, wówczas można użyć narzędzia Visualizer dotyczącego szeregów czasowych w celu graficznej prezentacji danych szeregów czasowych.

Aby używać narzędzia Visualizer dla szeregów czasowych:

1. Upewnij się, że wykonane zostały zadania integracji produktu IBM SPSS Modeler z produktem IBM InfoSphere Warehouse. Więcej informacji można znaleźć w temacie “Włączanie integracji z produktem IBM InfoSphere Warehouse” na stronie 50.
2. Kliknij dwukrotnie model surowy na palecie Modele.
3. Na karcie Serwer w oknie dialogowym kliknij przycisk Widok, aby wyświetlić narzędzie Visualizer w domyślnej przeglądarce WWW.

Modele użytkowe ISW Data Mining

Modele można tworzyć z węzłów ISW Decision Tree, ISW Association, ISW Sequence, ISW Regression i ISW Clustering zawartych w produkcie IBM SPSS Modeler.

Karta serwera modelu użytkowego ISW

Karta Serwer udostępnia opcje umożliwiające sprawdzanie spójności i uruchamianie narzędzia IBM Visualizer.

Produkt IBM SPSS Modeler może przeprowadzić sprawdzanie spójności poprzez zapisanie identycznego wygenerowanego klucza tekstowego zarówno w modelu IBM SPSS Modeler, jak i w modelu ISW. Sprawdzanie spójności jest wykonywane po kliknięciu przycisku **Sprawdź** na karcie Serwer. Więcej informacji można znaleźć w temacie “Zarządzanie modelami DB2” na stronie 54.

Narzędzie Visualizer jest jedyną metodą przeglądania modeli InfoSphere Warehouse Data Mining. Narzędzie może być opcjonalnie instalowane wraz z produktem InfoSphere Warehouse Data Mining. Więcej informacji można znaleźć w temacie “Włączanie integracji z produktem IBM InfoSphere Warehouse” na stronie 50.

- Kliknij opcję **Widok**, aby uruchomić narzędzie Visualizer. Treści, jakie zostaną wyświetlone w narzędziu, są zależne od typu wygenerowanego węzła. Na przykład narzędzie Visualizer zwróci widok klas przewidywanych, jeśli zostanie uruchomione z modelu użytkowego ISW Decision Tree.
- Kliknij opcję **Wyniki testu** (tylko drzewa decyzyjne i sekwencja), aby uruchomić narzędzie Visualizer i wyświetlić ogólną jakość wygenerowanego modelu.

Karta ustawień modelu użytkowego ISW

W produkcie IBM SPSS Modeler udostępniana jest zwykle tylko jedna predykcja i powiązane z nią prawdopodobieństwo lub współczynnik ufności. Ponadto dostępna dla użytkownika opcja, dzięki której możliwe jest wyświetlenie prawdopodobieństw dla każdego wyniku (co jest podobne do opcji dostępnej w regresji logistycznej), to opcja czasu oceny dostępna na karcie ustawień modelu użytkowego.

Dołącz ufności dla wszystkich klas. Wybranie tej opcji powoduje — w przypadku każdego z możliwych wyników dla zmiennej przewidywanej — dodanie kolumny zwracającej poziom ufności.

Karta podsumowania modelu użytkowego ISW

Karta Podsumowanie modelu użytkowego zawiera informacje na temat samego modelu (*Analiza*), zmiennych użytych w modelu (*Zmienne*), ustawień użytych podczas budowania modelu (*Ustawienia budowania*) i uczenia modelu (*Podsumowanie uczenia*).

Podczas przeglądania węzła po raz pierwszy karta Podsumowanie jest zwinięta. Aby zobaczyć wyniki będące przedmiotem zainteresowania, należy użyć rozszerzanego elementu sterującego po lewej stronie pozycji, aby ją rozwinąć, lub kliknąć przycisk **Rozwiń wszystko**, aby wyświetlić wszystkie wyniki. W celu ukrycia wyników po zakończeniu ich przeglądania należy użyć rozszerzanego elementu sterującego, aby zwinąć konkretne wyniki, jakie mają zostać ukryte, lub kliknąć przycisk **Zwiń wszystko**, aby zwinąć wszystkie wyniki.

Analiza. Wyświetla informacje na temat konkretnego modelu. Jeśli wykonano węzeł analizy dołączony do modelu użytkowego, wówczas informacje z tej analizy również pojawiają się w tej sekcji.

Zmienne. Na liście znajdują się zmienne użyte jako zmienne przewidywane i wejściowe podczas budowania modelu.

Ustawienia budowania. Zawiera informacje na temat ustawień użytych podczas budowania modelu.

Podsumowanie uczenia. Przedstawia typ modelu, strumień użyty do jego utworzenia, użytkownika, który go utworzył, informację, kiedy został utworzony, oraz czas, jaki był potrzebny do zbudowania modelu.

Przykłady dotyczące ISW Data Mining

Produkt IBM SPSS Modeler dla systemu Windows jest dostarczany z szeregiem strumieni demonstracyjnych, które ilustrują proces eksploracji bazy danych. Te strumienie można znaleźć w folderze instalacyjnym produktu IBM SPSS Modeler w katalogu:

```
\\Demos\Database_Modeling\IBM DB2 ISW
```

Uwaga: dostęp do folderu Demos można uzyskać z grupy programu IBM SPSS Modeler w menu Start systemu Windows.

Następujące strumienie mogą być używane razem sekwencyjnie jako przykład procesu eksploracji bazy danych:

- *1_upload_data.str* — służy do kasowania i wczytywania danych z pliku płaskiego do DB2.
- *2_explore_data.str* — używany jako przykład eksploracji danych za pomocą produktu IBM SPSS Modeler.
- *3_build_model.str* — używany do budowania modelu ISW Decision Tree.

- *4_evaluate_model.str* — używany jako przykład oceny modelu za pomocą produktu IBM SPSS Modeler.
- *5_deploy_model.str* — wdraża model na potrzeby oceniania w bazie danych.

Zestaw danych używany w strumieniach przykładowych dotyczy aplikacji kart kredytowych i prezentuje problem klasyfikacji z mieszaniną predyktorów jakościowych i ciągłych. Więcej informacji na temat tego zestawu danych zawiera następujący plik, który znajduje się w folderze instalacyjnym produktu IBM SPSS Modeler:

`\\Demos\Database_Modeling\IBM DB2 ISW\crx.names`

Ten zestaw danych jest dostępny z repozytorium UCI Machine Learning Repository pod adresem <http://archive.ics.uci.edu/ml/>.

Przykładowy strumień: Wczytywanie danych

Pierwszy przykładowy strumień o nazwie *1_upload_data.str* jest używany do kasowania i wczytywania danych z pliku płaskiego do programu DB2.

Węzeł wypełniania jest używany do obsługi wartości brakujących i zastępuje puste pola odczytywane z pliku tekstowego *crx.data* wartościami *NULL*.

Przykładowy strumień: Eksploracja danych

Drugi przykładowy strumień o nazwie *2_explore_data.str* jest używany do demonstrowania eksploracji danych w produkcie IBM SPSS Modeler.

Typowy krok używany podczas eksploracji danych obejmuje dołączenie węzła audytu danych do danych. Węzeł Audyt danych jest dostępny w palecie węzłów Wynik.

Dane wynikowe z węzła audytu danych można wykorzystać, aby uzyskać ogólny przegląd zmiennych i rozkładu danych. Dwukrotne kliknięcie wykresu w oknie Audyt danych generuje bardziej szczegółowy wykres, który pozwala na głębszą eksplorację pod kątem konkretnej zmiennej.

Przykładowy strumień: Budowa modelu

Trzeci przykładowy strumień — *3_build_model.str* — ilustruje budowanie modelu w produkcie IBM SPSS Modeler. Węzeł modelowania w bazie można dołączyć do strumienia, a następnie kliknąć dwukrotnie, aby określić ustawienia budowania.

Na kartach Model i Zaawansowany w węzle modelowania można regulować maksymalną głębokość drzewa, a ponadto można zatrzymać dalsze podziały węzła od momentu zbudowania początkowego drzewa decyzyjnego, ustawiając maksymalną czystość i minimalną liczbę obserwacji na węzeł wewnętrzny. Więcej informacji można znaleźć w temacie “ISW Decision Tree” na stronie 57.

Przykładowy strumień: Ocena modelu

Czwarty przykładowy strumień — *4_evaluate_model.str* — ilustruje zalety korzystania z produktu IBM SPSS Modeler na potrzeby modelowania w bazie danych. Po wykonaniu modelu można go dodać z powrotem do strumienia danych i ocenić model, używając kilku narzędzi oferowanych w produkcie IBM SPSS Modeler.

Po pierwszym otwarciu strumienia model użytkowy (*field16*) nie jest uwzględniany w strumieniu. Otwórz węzeł źródłowy CREDIT i upewnij się, że źródło danych zostało określone. Następnie — pod warunkiem że strumień *3_build_model.str* został uruchomiony w celu utworzenia modelu użytkowego *field16* w palecie Modele — możesz uruchomić rozłączone węzły, klikając przycisk **Wykonaj** na pasku narzędzi (przycisk z zielonym trójkątem). To spowoduje uruchomienie skryptu, który skopiuje model użytkowy *field16* do strumienia, połączy go z istniejącymi węzłami, a następnie uruchomi węzły końcowe w strumieniu.

Można również dołączyć węzeł analizy (dostępny na palecie wynikowej), aby utworzyć macierz zbieżności przedstawiającą wzorzec dopasowań między poszczególnymi wygenerowanymi zmiennymi (przewidywanymi) a odpowiadającymi im zmiennymi zależnymi. Uruchom węzeł analizy, aby zobaczyć wyniki.

Ponadto można utworzyć wykres korzyści, aby przedstawić postępy modelu w zakresie dokładności. Dołącz węzeł ewaluacji do wygenerowanego modelu, a następnie uruchom strumień, aby zobaczyć wyniki.

Przykładowy strumień: Wdrożenie modelu

Gdy dokładność modelu jest zadowalająca, można go wdrożyć, aby używać go z aplikacjami zewnętrznymi albo zapisać oceny z powrotem w bazie danych. W strumieniu przykładowym *5_deploy_model.str* dane są odczytywane z tabeli CREDIT. Gdy nastąpi uruchomienie węzła eksportu do bazy danych (o nazwie *deploy solution*), dane nie będą oceniane. Zamiast tego strumień utworzy publikowany plik graficzny *credit_scorer.pim* i publikowany plik parametrów *credit_scorer.par*.

Tak jak w poprzednim przykładzie strumień uruchomi skrypt, który skopiuje model użytkowy *field16* do strumienia z palety Modele, połączy go z istniejącymi węzłami, a następnie uruchomi węzły końcowe w strumieniu. W tym przypadku należy najpierw określić źródło danych w węzłach źródłowych i węzłach eksportu bazy danych.

Rozdział 6. Modelowanie w bazie za pomocą produktu IBM Netezza Analytics

IBM SPSS Modeler i IBM Netezza Analytics

Produkt IBM SPSS Modeler obsługuje integrację z produktem IBM Netezza Analytics, który zapewnia możliwość uruchamiania zaawansowanych analiz na serwerach IBM Netezza. Dostęp do tych funkcji można uzyskać za pośrednictwem graficznego interfejsu użytkownika produktu IBM SPSS Modeler i środowiska programistycznego zorientowanego na przepływ pracy, dzięki czemu można uruchamiać algorytmy eksploracji danych bezpośrednio w środowisku IBM Netezza.

Produkt IBM SPSS Modeler obsługuje integrację następujących algorytmów z produktu IBM Netezza Analytics:

- Drzewa decyzyjne
- K-średnie
- Sieć Bayesa
- Naive Bayes
- KNN
- Grupowanie dzielące
- PCA
- Drzewo regresji
- Regresja liniowa
- Szereg czasowy
- Uogólnione liniowe

Więcej informacji na temat tych algorytmów zawiera dokumentacja *IBM Netezza Analytics Developer's Guide* oraz *IBM Netezza Analytics Reference Guide*.

Wymagania dotyczące integracji z produktem IBM Netezza Analytics

Następujące warunki stanowią warunki wstępne do przeprowadzenia modelowania w bazie danych z użyciem produktu IBM Netezza Analytics. W celu upewnienia się, że te warunki są spełnione, konieczne mogą być konsultacje z administratorem bazy danych.

- Uruchomienie produktu IBM SPSS Modeler względem instalacji IBM SPSS Modeler Server na systemie Windows lub UNIX (z wyjątkiem zLinux, dla którego sterowniki ODBC IBM Netezza są niedostępne).
- Serwer IBM Netezza Performance Server, na którym działa pakiet IBM Netezza Analytics.

Uwaga: minimalna wymagana wersja serwera Netezza Performance Server (NPS) jest zależna od wymaganej wersji INZA i może być następująca:

- Dowolna wersja wyższa niż NPS 6.0.0 P8 będzie obsługiwać rozwiązanie INZA w wersjach wcześniejszych niż 2.0.
- W celu używania rozwiązania INZA 2.0 lub w wyższej wersji wymagana jest wersja NPS 6.0.5 P5 lub wyższa.

W celu zapewnienia funkcjonalności algorytmów Netezza Generalized Linear i Netezza Time Series wymagany jest produkt INZA 2.0 lub wyższe wersje. Wszystkie inne węzły Netezza w bazie danych wymagają wersji INZA 1.1 lub późniejszej.

- Źródło danych ODBC w celu połączenia z bazą danych IBM Netezza. Więcej informacji można znaleźć w temacie “Aktywacja integracji z produktem IBM Netezza Analytics” na stronie 76.
- Funkcje generowania i optymalizacji kodu SQL włączone w produkcie IBM SPSS Modeler. Więcej informacji można znaleźć w temacie “Aktywacja integracji z produktem IBM Netezza Analytics” na stronie 76.

Uwaga: Modelowanie w bazie danych i optymalizacja SQL wymagają włączenia na komputerze z programem IBM SPSS Modeler możliwości połączenia z serwerem IBM SPSS Modeler Server. Po włączeniu tej opcji można uzyskać dostęp do algorytmów baz danych, wstawić SQL do kolejki bezpośrednio z programu IBM SPSS Modeler i uzyskać dostęp do programu IBM SPSS Modeler Server. W celu sprawdzenia bieżącego statusu licencji należy wybrać z menu programu IBM SPSS Modeler następujące opcje.

Pomoc > Informacje o programie > Dodatkowe szczegóły

Po włączeniu możliwości połączenia na karcie Status licencji widoczna jest opcja **Aktywacja serwera**.

Aktywacja integracji z produktem IBM Netezza Analytics

Aktywacja integracji z produktem IBM Netezza Analytics obejmuje następujące etapy.

- Konfigurowanie produktu IBM Netezza Analytics
- Utworzenie źródła ODBC
- Aktywacja integracji w produkcie IBM SPSS Modeler
- Włączenie opcji generowania i optymalizacji kodu SQL w IBM SPSS Modeler

Te etapy zostały opisane w kolejnych sekcjach.

Konfigurowanie produktu IBM Netezza Analytics

Informacje dotyczące instalowania i konfigurowania produktu IBM Netezza Analytics zawiera dokumentacja produktu IBM Netezza Analytics — przede wszystkim *IBM Netezza Analytics Installation Guide*. W tym podręczniku sekcja *Setting Database Permissions* zawiera szczegółowe informacje dotyczące skryptów, które należy uruchomić, aby zezwolić strumieniom IBM SPSS Modeler na zapisywanie do bazy danych.

Uwaga: jeśli używane będą węzły, które polegają na obliczeniu macierzy (Netezza PCA i Netezza Linear Regression), wówczas silnik Netezza Matrix Engine należy zainicjować poprzez uruchomienie metody `CALL NZM..INITIALIZE();` w przeciwnym wypadku wykonanie składowanych nie powiedzie się. Inicjalizacja jest jednorazowym procesem konfiguracji w przypadku każdej bazy danych.

Tworzenie źródła ODBC dla produktu IBM Netezza Analytics

Aby umożliwić połączenie bazy danych IBM Netezza i produktu IBM SPSS Modeler, należy utworzyć nazwę źródła danych (DSN) ODBC.

Przed utworzeniem DSN wymagana jest podstawowa znajomość źródeł danych i sterowników ODBC, a także obsługa bazy danych w produkcie IBM SPSS Modeler.

W przypadku uruchamiania w trybie rozproszonym w produkcie IBM SPSS Modeler Server należy utworzyć DSN na komputerze serwera. W przypadku uruchamiania w trybie lokalnym (klienta) należy utworzyć DSN na komputerze klienckim.

Klienci Windows

1. Z dysku CD *Netezza Client* uruchom plik *nzodbcsetup.exe*, aby uruchomić instalator. W celu zainstalowania sterownika postępuj zgodnie z instrukcjami wyświetlanymi na ekranie. Pełne instrukcje zawiera podręcznik instalacji i konfiguracji sterowników ODBC, JDBC oraz OLE DB dla IBM Netezza.

a. Utwórz DSN.

Uwaga: kolejność opcji w menu jest zależna od wersji systemu Windows.

- **Windows XP.** Z menu Start wybierz opcję **Panel sterowania**. Kliknij dwukrotnie opcję **Narzędzia administracyjne**, a następnie kliknij dwukrotnie opcję **Źródła danych (ODBC)**.
- **Windows Vista.** Z menu Start wybierz opcję **Panel sterowania**, a następnie **Konserwacja systemu**. Kliknij dwukrotnie pozycję **Narzędzia administracyjne**, wybierz pozycję **Źródła danych (ODBC)**, a następnie kliknij opcję **Otwórz**.

- **Windows 7.** Z menu Start wybierz pozycję **Panel sterowania**, następnie **System i zabezpieczenia**, a następnie **Narzędzia administracyjne**. Wybierz pozycję **Źródła danych (ODBC)**, a następnie kliknij opcję **Otwórz**.
- b. Kliknij kartę **Systemowe źródło danych DSN** i kliknij opcję **Dodaj**.
2. Wybierz z listy opcję **NetezzaSQL** i kliknij przycisk **Finish** (Zakończ).
 3. Na karcie **DSN Options** (Opcje DSN) na ekranie konfiguracji sterownika ODBC Netezza wpisz wybraną nazwę źródła danych, nazwę hosta lub adres IP serwera IBM Netezza, numer portu dla połączeń, nazwę bazy danych używanej instancji IBM Netezza, a także własną nazwę użytkownika i hasło na potrzeby połączenia z bazą danych. Aby uzyskać wyjaśnienie pól, kliknij przycisk **Help** (Pomoc).
 4. Kliknij przycisk **Test Connection** (Testuj połączenie) i upewnij się, że możliwe jest nawiązanie połączenia z bazą danych.
 5. Po pomyślnym nawiązaniu połączenia kliknij kilkakrotnie przycisk **OK**, aby zamknąć ekran administratora źródła danych ODBC.

Serwery Windows

Procedura dla serwera Windows jest taka sama, jak procedura dotycząca klienta dla systemu Windows XP.

Serwery UNIX lub Linux

Poniższa procedura dotyczy serwerów UNIX lub Linux (z wyjątkiem zLinux, dla którego sterowniki ODBC IBM Netezza są niedostępne).

1. Z dysku CD/DVD Netezza Client skopiuj odpowiedni plik `<platforma>cli.package.tar.gz` do lokalizacji tymczasowej na serwerze.
2. Wyodrębnij zawartość archiwum, stosując polecenia **gunzip** i **untar**.
3. Dodaj uprawnienia do wykonywania skryptu *unpack*, który został wyodrębniony.
4. Uruchom skrypt, reagując na monity wyświetlane na ekranie.
5. Przeprowadź edycję pliku *modelersrv.sh* w taki sposób, aby plik zawierał poniższe wiersze.

```
. <SDAP Install Path>/odbc.sh
LD_LIBRARY_PATH_64=$LD_LIBRARY_PATH:/usr/local/nz/lib64; export LD_LIBRARY_PATH_64
NZ_ODBC_INI_PATH=<SDAP Install Path>; export NZ_ODBC_INI_PATH
```

Na przykład:

```
./usr/IBM/SPSS/SDAP/odbc.sh
LD_LIBRARY_PATH_64=$LD_LIBRARY_PATH:/usr/local/nz/lib64; export LD_LIBRARY_PATH_64
NZ_ODBC_INI_PATH=/usr/IBM/SPSS/SDAP; export NZ_ODBC_INI_PATH
```

6. Odszukaj plik `/usr/local/nz/lib64/odbc.ini` i skopiuj jego zawartość do pliku `odbc.ini`, który jest instalowany przy użyciu sterownika SDAP (tego, który jest zdefiniowany przez zmienną środowiskową \$ODBCINI).

Uwaga: w przypadku 64-bitowych systemów Linux parametr **Driver** niepoprawnie odwołuje się do sterownika 32-bitowego. Podczas kopiowania zawartości `odbc.ini` w poprzednim kroku należy przeprowadzić odpowiednią edycję ścieżki w tym parametrze, na przykład:

```
/usr/local/nz/lib64/libnzodbc.so
```

7. Przeprowadź edycję parametrów w definicji DSN Netezza w taki sposób, aby odzwierciedlały używaną bazę danych.
8. Zrestartuj produkt IBM SPSS Modeler Server i przetestuj używanie kodów eksploracji w bazie danych Netezza na kliencie.

Aktywacja integracji z produktem IBM Netezza Analytics w produkcji IBM SPSS Modeler

1. Z menu głównego produktu IBM SPSS Modeler wybierz pozycję **Narzędzia > Opcje > Aplikacje pomocnicze**.

2. Kliknij kartę **IBM Netezza**.

Włącz integrację Netezza Data Mining. Włącza paletę Modelowanie w bazie (jeśli nie jest jeszcze wyświetlana) u dołu okna IBM SPSS Modeler i dodaje węzły przeznaczone dla algorytmów Netezza Data Mining.

Połączenie Netezza. Kliknij przycisk **Edytuj** i wybierz łańcuch połączenia Netezza, który został wcześniej skonfigurowany podczas tworzenia źródła ODBC. Więcej informacji można znaleźć w temacie “Tworzenie źródła ODBC dla produktu IBM Netezza Analytics” na stronie 76.

Włączenie opcji generowania i optymalizacji kodu SQL

Istnieje możliwość pracy z bardzo dużymi zbiorami danych, dlatego ze względu na wydajność należy włączyć opcje generowania i optymalizacji kodu SQL w produkcie IBM SPSS Modeler.

1. Z menu programu IBM SPSS Modeler wybierz:
Narzędzia > Właściwości strumienia > Opcje
2. Kliknij opcję **Optymalizacja** w panelu nawigacji.
3. Upewnij się, że włączona jest opcja **Generuj kod SQL kierowany do bazy**. To ustawienie jest niezbędne, ponieważ zapewnia poprawne działanie modelowania w bazie danych.
4. Wybierz opcje **Optymalizuj operacje generujące kod SQL** i **Optymalizuj inne wykonywane operacje** (nie jest to ściśle wymagane, ale zdecydowanie zalecane w celu poprawy wydajności).

Budowanie modeli za pomocą produktu IBM Netezza Analytics

Dla każdego obsługiwane algorytmu istnieje odpowiadający mu węzeł modelowania. Dostęp do węzłów modelowania IBM Netezza można uzyskać z karty Modelowanie w bazie na palecie węzłów.

Zagadnienia dotyczące danych

Pola w źródle danych zawierają zmienne różnych typów danych, co jest uzależnione od węzła modelowania. W produkcie IBM SPSS Modeler typy danych są określane mianem *poziomy pomiaru*. W karcie Zmienne dla węzła modelowania używane są ikony oznaczające dozwolone typy poziomów pomiaru dla zmiennych wejściowych i przewidywanych.

Zmienna przewidywana Zmienna przewidywana to zmienna, której wartość próbujemy przewidzieć. Gdy możliwe jest określenie zmiennej przewidywanej, wówczas tylko jedno pole danych źródłowych można wybrać jako zmienną przewidywaną.

Zmienna identyfikacyjna rekordu Określa zmienną używaną w celu unikalnej identyfikacji każdej obserwacji. Na przykład może to być zmienna identyfikacyjna, taka jak *CustomerID*. Jeśli dane źródłowe nie zawierają zmiennej identyfikacyjnej, można utworzyć tę zmienną za pomocą węzła wyliczeń, co przedstawia poniższa procedura.

1. Wybierz węzeł źródłowy.
2. Na karcie Zmienne w palecie węzłów kliknij dwukrotnie węzeł wyliczeń.
3. Otwórz węzeł wyliczeń, klikając dwukrotnie jego ikonę na obszarze roboczym.
4. W polu **Zmienna wyliczana** wpisz (na przykład) ID.
5. W polu **Formuła** wpisz @INDEX i kliknij przycisk **OK**.
6. Połącz węzeł wyliczeń z pozostałą częścią strumienia.

Uwaga: Jeśli z bazy danych Netezza pobierane są długie dane liczbowe przy użyciu typu danych NUMERIC(18,0), wówczas produkt SPSS Modeler może czasami zaokrąglać dane w górę podczas importu. Aby uniknąć tego problemu, należy zapisywać dane jako dane typu BIGINT lub NUMERIC(36,0).

Uwaga: Z powodu ograniczeń typów zmiennych, które mogą być używane, zmienna o beztypowym poziomie pomiaru i roli ID rekordu nie pojawia się w węzle modelowania w bazie danych Netezza (np. K-średnich).

Postępowanie z wartościami null

Jeśli wartości wejściowe zawierają wartości null, stosowanie niektórych węzłów Netezza może spowodować komunikaty o błędach lub długo działające strumienie, dlatego zalecamy usunięcie rekordów zawierających wartości null. Należy użyć poniższej metody.

1. Dołącz węzeł selekcji do węzła źródłowego.
2. Ustaw opcję **Tryb** węzła selekcji na wartość **Odrzuć**.
3. Wprowadź poniższy łańcuch do pola **Warunek**:
`@NULL(field1) [or @NULL(field2)[... or @NULL(fieldN)]`

Upewnij się, że uwzględniona została każda zmienna wejściowa.

4. Połącz węzeł selekcji z pozostałą częścią strumienia.

Wyniki modelu

Strumień zawierający węzeł modelowania Netezza może zwracać trochę inne wyniki za każdym razem, gdy jest uruchamiany. Dzieje się tak, ponieważ kolejność, w jakiej węzeł odczytuje źródło danych, nie jest zawsze taka sama, gdyż przed budowaniem modelu dane są wczytywane do tabel tymczasowych. Jednak różnice wywołane przez ten efekt są pomijalne.

Komentarze ogólne

- W produkcie IBM SPSS Collaboration and Deployment Services nie jest możliwe ocenianie konfiguracji przy użyciu strumieni zawierających węzły modelowania w bazie danych IBM Netezza.
- W przypadku modeli utworzonych przez węzły Netezza nie jest możliwy eksport ani import kodu PMML.

Modele Netezza — opcje zmiennych

Na karcie Zmienne można zdecydować, czy mają zostać użyte ustawienia roli zmiennej już zdefiniowane w węzłach poprzedzających, czy też przypisania zmiennych mają zostać dokonane ręcznie.

Użyj wstępnie zdefiniowanych ról. Ta opcja korzysta z ustawień roli (zmienne przewidywane, predyktory itd.) z poprzedzającego węzła Typ (lub z karty Typy poprzedzającego węzła źródłowego).

Użyj niestandardowych przypisań. Opcję tę należy wybrać w celu ręcznego przypisania zmiennych przewidywanych, predyktorów oraz innych ról na tym ekranie.

Zmienne. Aby ręcznie przypisać pozycje z tej listy do różnych zmiennych ról po prawej stronie ekranu, należy użyć klawiszy strzałek. Ikony wskazują prawidłowe poziomy pomiaru dla każdego pola roli.

Kliknij przycisk **Wszystkie**, aby wybrać wszystkie pola na liście, lub kliknij przycisk poziomego pojedynczego pomiaru, aby wybrać wszystkie pola, które posiadają ten poziom pomiaru.

Zmienna przewidywana. Wybierz jedną zmienną jako zmienną przewidywaną dla predykcji. W przypadku algorytmu Uogólnione modele liniowe wybierz także zmienną **Próby** na tym ekranie.

Identyfikator rekordu. Zmienna, która będzie używana jako unikatowy identyfikator rekordu.

Predyktory (dane wejściowe). Wybierz jedną lub więcej zmiennych jako dane wejściowe dla predykcji.

Modele Netezza — opcje serwera

Na karcie Serwer określ bazę danych IBM Netezza, w której zostanie zbudowany model.

Szczegóły serwera DB Netezza. W tym miejscu należy określić szczegóły połączenia z bazą danych, która będzie używana na potrzeby modelu.

- **Użyj połączenia zdefiniowanego w strumieniu.** (ustawienie domyślne) Wybranie tej opcji powoduje użycie szczegółów połączenia określonych w węźle poprzedzającym, na przykład węzle źródła bazy danych. *Uwaga:* ta opcja działa tylko wówczas, gdy wszystkie węzły poprzedzające mogą wykorzystywać funkcję analizy wstępnej SQL. W takim przypadku nie ma potrzeby przesuwać danych poza bazę danych, ponieważ SQL w pełni implementuje wszystkie węzły poprzedzające.
- **Przenieś dane do bazy przez połączenie.** Przenosi dane do bazy danych określonej w tym miejscu. Dzięki temu modelowanie może działać, jeśli dane znajdują się w innej bazie danych IBM Netezza lub baza pochodzi od innego dostawcy, a także w przypadku, gdy dane znajdują się w pliku płaskim. Ponadto dane są przenoszone z powrotem do bazy danych określonej w tej opcji, jeśli dane zostały wyodrębnione, ponieważ węzeł nie zrealizował analizy wstępnej SQL. Kliknij przycisk **Edycja**, aby przeglądać w poszukiwaniu połączenia i wybrać połączenie.
Ostrzeżenie: produkt IBM Netezza Analytics jest zwykle używany z bardzo dużymi zestawami danych. Przesyłanie dużych ilości danych między bazami danych lub z bazy i do bazy może być bardzo czasochłonne i w miarę możliwości należy tego unikać.

Uwaga: Nazwa źródła danych ODBC jest osadzona w każdym strumieniu IBM SPSS Modeler. Jeśli strumień utworzony w jednym hoście zostanie wykonany na innym hoście, wówczas nazwa źródła danych musi być taka sama na każdym hoście. Alternatywnie inne źródło danych można wybrać na karcie Serwer w każdym węźle źródłowym lub węźle modelowania.

Modele Netezza — opcje modelu

Na karcie Opcje modelu można zdecydować o wyborze nazwy dla modelu lub o jej wygenerowaniu automatycznie. Dla opcji oceniania można ustawić wartości domyślne.

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu, w przypadkach gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Zastąp istniejący, jeśli nazwa została użyta. Jeśli to pole wyboru zostanie zaznaczone, wówczas wszelkie istniejące modele o tej samej nazwie zostaną zastąpione.

Udostępnij do oceniania. W tym miejscu można ustawić wartości domyślne dla opcji oceniania, które będą widoczne w oknie dialogowym dla modelu użytkowego. Szczegółowe informacje o opcjach zawiera temat pomocy dla karty Ustawienia dla konkretnego modelu użytkowego.

Zarządzanie modelami Netezza

Budowanie modelu IBM Netezza za pośrednictwem produktu IBM SPSS Modeler powoduje utworzenie modelu w produkcie IBM SPSS Modeler oraz utworzenie lub zastąpienie modelu w bazie danych Netezza. Model tego rodzaju w produkcie IBM SPSS Modeler odwołuje się do zawartości modelu bazy danych zapisanego na serwerze bazy danych. Produkt IBM SPSS Modeler może przeprowadzić sprawdzanie spójności poprzez zapisanie identycznego wygenerowanego klucza tekstowego zarówno w modelu IBM SPSS Modeler, jak i w modelu Netezza.

Nazwa dla każdego modelu Netezza jest wyświetlana w kolumnie *Informacje o modelu* w oknie dialogowym Wymień modele bazy danych. Nazwa modelu dla modelu IBM SPSS Modeler jest wyświetlana jako klucz modelu na karcie Serwer modelu IBM SPSS Modeler (po umieszczeniu w strumieniu).

Przycisku Sprawdź można użyć, aby sprawdzić, czy klucze modelu w modelu IBM SPSS Modeler i modelu Netezza są zgodne. Jeśli w Netezza nie można znaleźć żadnego modelu o tej samej nazwie lub jeśli klucze modelu są niezgodne, wówczas oznacza to, że model Netezza został usunięty lub ponownie zbudowany po zbudowaniu modelu IBM SPSS Modeler.

Lista modeli bazy danych

W produkcie IBM SPSS Modeler dostępne jest okno dialogowe, w którym można wyświetlić listę modeli zapisanych w produkcie IBM Netezza. W tym oknie modele można również usuwać. Dostęp do tego okna można uzyskać z okna

dialogowego IBM Helper Applications, a także z okien dialogowych do budowania, przeglądania i okna dialogowego Zastosuj dla węzłów związanych z eksploracją danych w IBM Netezza. Dla każdego modelu wyświetlane są następujące informacje:

- Nazwa modelu (używana w celu sortowania listy).
- Nazwa właściciela.
- Algorytm użyty w modelu.
- Bieżący stan modelu; na przykład Zakończono.
- Data utworzenia modelu.

Netezza Regression Tree

Drzewo regresji to algorytm bazujący na drzewie, który wielokrotnie dzieli próbę obserwacji, aby uzyskać podzbiór tego samego rodzaju odpowiednio do wartości liczbowej zmiennej przewidywanej. Drzewa regresji, podobnie jak drzewa decyzyjne, dekomponują dane na podzbiory, w których liście drzewa odpowiadają wystarczająco małym lub wystarczająco jednostajnym podzbiорom. Podziały są wybierane w celu zmniejszenia rozproszenia przewidywanych wartości atrybutu, dzięki czemu te wartości mogą być odpowiednio dobrze przewidziane na podstawie wartości średnich na liściach.

Opcje budowania drzewa regresji Netezza — wzrost drzewa

Dla wzrostu drzewa i przycinania drzewa można ustawić opcje budowania.

Dla wzrostu drzewa dostępne są następujące opcje budowania:

Maksymalna głębokość drzewa. Maksymalna liczba liści, do której może rosnać drzewo poniżej węzła głównego, czyli jest to liczba rekurencyjnych podziałów próby. Wartością domyślną jest 62, co jest maksymalną głębokością drzewa na potrzeby modelowania.

Uwaga: Jeśli przeglądarka w modelu użytkowym przedstawia tekstową reprezentację modelu, wówczas wyświetlanych jest maksymalnie 12 poziomów drzewa.

Kryteria podziału. Te opcje kontrolują miejsce, w którym następuje zatrzymanie podziału drzewa. Jeśli nie zamierzasz używać wartości domyślnych, kliknij opcję **Dostosuj**, a następnie zmień wartości.

- **Ewaluacyjna miara podziału.** Miara ewaluacyjna klasy ocenia najlepsze miejsce podziału drzewa.

Uwaga: Aktualnie jedyną możliwą opcją jest wariancja.

- **Minimalne ulepszenie dla podziałów.** Minimalna wartość, o jaką musi zostać zmniejszone zanieczyszczenie, zanim w drzewie zostanie utworzony nowy podział. Celem budowy drzewa jest utworzenie podgrup o podobnych wartościach wyjściowych — czyli minimalizowanie zanieczyszczeń w ramach każdego węzła. Jeśli najlepszy podział dla gałęzi zmniejsza zanieczyszczenia o mniej, niż określono w kryteriach podziału, wówczas gałąź nie jest dzielona.
- **Minimalna liczba instancji dla podziału.** Minimalna liczba rekordów, jaka może zostać podzielona. Jeśli liczba pozostałych niepodzielonych rekordów jest mniejsza niż ta liczba, wówczas dalsze podziały nie są wykonywane. Tego pola można użyć, aby zapobiec tworzeniu małych podgrup w drzewie.

Statystyki. Ten parametr definiuje liczbę statystyk uwzględnionych w modelu. Wybierz jedną z poniższych opcji:

- **Wszystko.** Uwzględnione są wszystkie statystyki dotyczące kolumny i wszystkie statystyki dotyczące wartości.

Uwaga: Ten parametr zawiera maksymalną liczbę statystyk i może wpłynąć na wydajność systemu. Jeśli wyświetlanie modelu w formacie graficznym nie jest wymagane, należy określić wartość **Brak**.

- **Kolumny.** Uwzględnione są statystyki dotyczące kolumny.
- **Brak.** Uwzględnione są tylko statystyki, które są wymagane do oceny modelu.

Opcje budowania drzewa regresji Netezza — przycinanie drzewa

Za pomocą opcji przycinania można określić kryteria przycinania dla drzewa regresji. Celem przycinania jest zmniejszenie ryzyka przeuczenia poprzez usunięcie przerośniętych podgrup, które nie poprawiają oczekiwanej dokładności w przypadku nowych danych.

Miara przycinania. Miara przycinania zapewnia, że oszacowana dokładność modelu mieści się w dozwolonych limitach po usunięciu liścia z drzewa. Wybierz jedną z poniższych miar.

- **mse.** Średni błąd kwadratowy — (domyślnie) mierzy bliskość dopasowanej linii do punktów danych.
- **r².** R-kwadrat — mierzy proporcję zmienności w zmiennej zależnej wyjaśnionej przez model regresji.
- **Pearsona.** Współczynnik korelacji Pearsona — mierzy siłę zależności między zmiennymi zależnymi liniowo, które mają rozkład normalny.
- **Spearman.** Współczynnik korelacji Spearmana — wykrywa zależności nieliniowe, które wydają się słabe w przypadku uwzględnienia korelacji Pearsona, ale w rzeczywistości mogą być silne.

Dane dla przycinania. W celu oszacowania oczekiwanej dokładności nowych danych można użyć części lub wszystkich danych uczących. Alternatywnie do tego celu można użyć osobnego zestawu danych do przycinania z podanej tabeli.

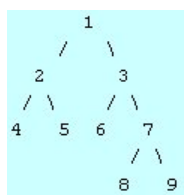
- **Użyj wszystkich danych uczących.** Ta opcja (domyślna) powoduje, że wykorzystywane są wszystkie dane uczące w celu oszacowania dokładności modelu.
- **Użyj % danych uczących do przycinania.** Ta opcja umożliwia podział danych na dwa zestawy — jeden jest używany do nauki, a drugi do przycinania, przy czym określony w opcji procent dotyczy danych do przycinania. Jeśli chcesz określić wartość początkową generatora liczb losowych, aby upewnić się, że dane zostaną podzielone tak samo za każdym razem, gdy uruchomisz strumień, wybierz opcję **Replikuj wyniki**. Możesz określić wartość całkowitą w polu **Wartość startowa generatora użyta do przycinania** lub kliknąć opcję **Utwórz**, co spowoduje utworzenie pseudolosowej wartości całkowitej.
- **Użyj danych z istniejącej tabeli.** Określ nazwę tabeli osobnego zestawu danych do przycinania w celu oszacowania dokładności modelu. Takie postępowanie jest traktowane jako bardziej niezawodne niż korzystanie z danych uczących. Jednak użycie tej opcji może spowodować usunięcie dużego zestawu danych z zestawu uczącego, co spowoduje obniżenie jakości drzewa decyzyjnego.

Netezza Divisive Clustering

Grupowanie dzielące to metoda analizy skupień, w której algorytm jest uruchamiany wielokrotnie w celu podziału skupień na podgrupy, aż do osiągnięcia określonego punktu zatrzymania.

Tworzenie skupień zaczyna się od pojedynczego skupienia zawierającego wszystkie instancje uczące (rekordy). Pierwsza iteracja algorytmu dzieli zestaw danych na dwie podgrupy, a każda kolejna iteracja dzieli je na dalsze podgrupy. Kryteria zatrzymania są określone jako maksymalna liczba iteracji, maksymalna liczba poziomów, na które podzielone są dane, a także minimalna wymagana liczba instancji do dalszego podziału.

Wynikowe drzewo hierarchiczne może służyć do klasyfikowania instancji poprzez propagowanie ich w dół od skupienia głównego, jak w poniższym przykładzie.



Rysunek 4. Przykład drzewa grupowania dzielącego

Na każdym poziomie wybierana jest najlepiej dopasowana podgrupa z uwzględnieniem odległości instancji od środków podgrup.

Gdy instancje są oceniane po zastosowaniu poziomu hierarchii -1 (domyślnie), wówczas ocena zwraca tylko jedno skupienie-liść, ponieważ liście są oznaczone liczbami ujemnymi. W tym przykładzie byłoby to jedno ze skupień 4, 5, 6, 8 lub 9. Jeśli jednak poziom hierarchii zostanie ustawiony na przykład na 2, wówczas ocena zwróci jedno ze skupień na poziomie drugim poniżej skupienia głównego, czyli 4, 5, 6 lub 7.

Opcje zmiennej grupowania dzielącego Netezza

Na karcie Zmienne można zdecydować, czy mają zostać użyte ustawienia roli zmiennej już zdefiniowane w węzłach poprzedzających, czy też przypisania zmiennych mają zostać dokonane ręcznie.

Użyj wstępnie zdefiniowanych ról. Ta opcja korzysta z ustawień roli (zmienne przewidywane, predyktory itd.) z poprzedzającego węzła Typ (lub z karty Typy poprzedzającego węzła źródłowego).

Użyj niestandardowych przypisań. Opcję tę należy wybrać w celu ręcznego przypisania zmiennych przewidywanych, predyktorów oraz innych ról na tym ekranie.

Zmienne. Aby ręcznie przypisać pozycje z tej listy do różnych zmiennych ról po prawej stronie ekranu, należy użyć klawiszy strzałek. Ikony wskazują prawidłowe poziomy pomiaru dla każdego pola roli.

Kliknij przycisk **Wszystkie**, aby wybrać wszystkie pola na liście, lub kliknij przycisk poziomu pojedynczego pomiaru, aby wybrać wszystkie pola, które posiadają ten poziom pomiaru.

Identyfikator rekordu. Zmienna, która będzie używana jako unikatowy identyfikator rekordu.

Predyktory (dane wejściowe). Wybierz jedną lub więcej zmiennych jako dane wejściowe dla predykcji.

Opcje budowania grupowania dzielącego Netezza

Na karcie Opcje budowania można ustawić wszystkie opcje budowania modelu. Można oczywiście po prostu kliknąć przycisk **Uruchom** w celu zbudowania modelu ze wszystkimi opcjami domyślnymi, lecz w normalnej sytuacji użytkownik potrzebuje zwykle dostosować proces budowy do swoich celów.

Miara odległości. Metoda używana w celu pomiaru odległości między punktami danych; większa odległość oznacza większe niepodobieństwo. Opcje są następujące:

- **Euklidesowa.** (opcja domyślna) Odległość między dwoma punktami wyliczona jako długość łączącej je linii prostej.
- **Manhattan.** Odległość między dwoma punktami jest sumą bezwzględnych różnic między ich współrzędnymi.
- **Canberra.** Podobna do odległości Manhattan, ale bardziej czuła w przypadku punktów danych położonych bliżej początku układu współrzędnych.
- **Maksimum.** Odległość między dwoma punktami jest obliczana jako największa z różnic wzdłuż dowolnej współrzędnej.

Maksymalna liczba iteracji. Algorytm działa poprzez wykonywanie kilku iteracji tego samego procesu. Ta opcja umożliwia zatrzymanie uczenia modelu po upływie podanej liczby iteracji.

Maksymalna głębokość drzew skupień. Maksymalna liczba poziomów, na które zostanie podzielony zbiór danych.

Replikacja wyników. Zaznacz to pole wyboru, jeśli zamierzasz ustawić wartość początkową generatora liczb losowych, co umożliwi replikowanie wyników. Możesz określić wartość całkowitą lub kliknąć opcję **Utwórz**, co spowoduje utworzenie pseudolosowej wartości całkowitej.

Minimalna liczba instancji dla podziału. Minimalna liczba rekordów, jaka może zostać podzielona. Jeśli liczba pozostałych niepodzielonych rekordów jest mniejsza niż ta liczba, wówczas dalsze podziały nie są wykonywane. Tego pola można użyć, aby zapobiec tworzeniu bardzo małych podgrup w drzewie skupień.

Netezza Generalized Linear

Regresja liniowa to stosowana od dawna technika statystyczna przeznaczona do klasyfikowania rekordów na podstawie wartości numerycznych zmiennych wejściowych. Regresja liniowa dopasowuje prostą linię lub powierzchnię, która minimalizuje rozbieżności między wartościami przewidywanymi a rzeczywistymi wynikami. Modele liniowe są użyteczne w modelowaniu szerokiej gamy rzeczywistych zjawisk, co jest możliwe dzięki prostocie zarówno pod względem uczenia, jak i zastosowań modelu. Jednak modele liniowe zakładają rozkład normalny w zależnej zmiennej (przewidywanej) oraz liniowy wpływ zmiennych niezależnych (predyktorów) na zmienną zależną.

Istnieje wiele sytuacji, w których regresja liniowa jest użyteczna, ale powyższe założenie nie obowiązuje. Na przykład podczas modelowania wyboru klienta między dyskretną liczbą produktów zmienna zależna może przyjmować rozkład wielomianowy. I podobnie podczas modelowania dochodów w kontekście wieku — dochody zwykle zwiększają się wraz z wiekiem, ale połączenie między tymi zmiennymi niekoniecznie musi być tak oczywiste jak linia prosta.

W takich sytuacjach można użyć uogólnionego modelu liniowego. Uogólnione modele liniowe rozszerzają model regresji liniowej w taki sposób, że zmienna zależna jest powiązana ze zmiennymi predyktora przez określoną funkcję łączenia, którą można wybrać spośród odpowiednich funkcji. Model pozwala ponadto, aby zmienna zależna nie miała rozkładu normalnego, a miała na przykład rozkład Poissona.

Algorytm iteracyjnie poszukuje najlepiej dopasowanego modelu, aż do określonej liczby iteracji. Podczas obliczania najlepszego dopasowania błąd jest reprezentowany przez sumę kwadratów różnic między wartością przewidywaną a rzeczywistą zmienną zależną.

Opcje zmiennej modelowania z użyciem uogólnionego modelu liniowego Netezza

Na karcie Zmienne można zdecydować, czy mają zostać użyte ustawienia roli zmiennej już zdefiniowane w węzłach poprzedzających, czy też przypisania zmiennych mają zostać dokonane ręcznie.

Użyj wstępnie zdefiniowanych ról. W przypadku tej opcji używane są ustawienia roli (np. zmienne przewidywane lub predyktory) z poprzedzającego węzła Typ (albo z karty Typy poprzedzającego węzła źródłowego).

Użyj niestandardowych przypisań. Opcję tę należy wybrać w celu ręcznego przypisania zmiennych przewidywanych, predyktorów oraz innych ról na tym ekranie.

Zmienne. Aby ręcznie przypisać pozycje z tej listy do różnych zmiennych ról po prawej stronie ekranu, należy użyć klawiszy strzałek. Ikony wskazują prawidłowy poziomy pomiar dla każdego pola roli.

Kliknij przycisk **Wszystkie**, aby wybrać wszystkie pola na liście, lub kliknij przycisk poziomu pojedynczego pomiaru, aby wybrać wszystkie pola, które posiadają ten poziom pomiaru.

Zmienna przewidywana. Wybierz jedną zmienną jako zmienną przewidywaną dla predykcji.

Identyfikator rekordu. Zmienna, która będzie używana jako unikatowy identyfikator rekordu. Wartości tej zmiennej muszą być unikalne dla każdego rekordu, np. muszą to być identyfikatory klientów.

Waga instancji. Określ zmienną, w której używane będą wagi instancji. Waga instancji to waga na wiersz danych wejściowych. Domyślnie obowiązuje założenie, że wszystkie rekordy wejściowe mają równą wagę względną. Wagę można zmienić, przypisując poszczególne wagi do rekordów wejściowych. Zmienna, którą określa użytkownik, musi zawierać wagę liczbową dla każdego wiersza danych wejściowych.

Predyktory (dane wejściowe). Umożliwia wybór zmiennej wejściowej lub kilku zmiennych. To działanie jest podobne jak w przypadku ustawienia roli zmiennej na *Zmienna wejściowa* w węzle Typ.

Opcje zmiennej modelowania z użyciem uogólnionego modelu liniowego Netezza — informacje ogólne

Na karcie Opcje modelu można zdecydować o wyborze nazwy dla modelu lub o jej wygenerowaniu automatycznie. Możliwe jest definiowanie różnych ustawień dotyczących modelu, funkcji łączenia, interakcji zmiennej wejściowej (jeśli istnieją), a także ustawianie wartości domyślnych dla opcji oceniania.

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu, w przypadkach gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Opcje zmiennych. Możesz określać role zmiennych wejściowych na potrzeby budowania modelu.

Ustawienia ogólne. Te ustawienia dotyczą kryteriów zatrzymania dla algorytmu.

- **Maksymalna liczba iteracji.** Maksymalna liczba iteracji, jakie wykona algorytm; minimum to 1, maksimum wynosi 20.
- **Błąd maksymalny (1e).** Wartość błędu maksymalnego (w notacji matematycznej), przy którym algorytm powinien zatrzymać wyszukiwanie modelu najlepiej dopasowanego. Minimum to 0, wartością domyślną jest -3, średnia to 1E-3 lub 0,001.
- **Próg wartości błędu nieznacznego (1e).** Wartość (w notacji matematycznej), poniżej której błędy są traktowane jako posiadające wartość zero. Minimum to -1, wartością domyślną jest -7, co oznacza, że wartości błędów poniżej 1E-7 (lub 0,0000001) są traktowane jako nieznaczące.

Ustawienia rozkładu. Te ustawienia dotyczą rozkładu zależnej zmiennej (przewidywanej).

- **Rozkład zmiennej odpowiedzi.** Rodzaj rozkładu; jeden z następujących: **Bernoulliego** (domyślny), **Gausa**, **Poissona**, **Dwumianowy**, **Ujemny dwumianowy**, **Walda** (Odwrócony Gaussa) oraz **Gamma**.
- **Parametry.** (Dotyczy tylko rozkładu Poissona i dwumianowego) W polu **Określ parametry** należy określić co najmniej jedną z poniższych opcji:
 - W celu uzyskania automatycznego oszacowania parametru na podstawie danych wybierz opcję **Domyślne**.
 - Aby pozwolić na optymalizację quasi-wiarygodności rozkładu, wybierz opcję **Quasi**.
 - Aby jawnie określić wartość parametru, wybierz opcję **Jawnie**.

(Dotyczy tylko rozkładu dwumianowego) Określ kolumnę tabeli wejściowej, która będzie używana jako zmienna prób wymagana w przypadku rozkładu dwumianowego. Ta kolumna zawiera liczbę prób dla rozkładu dwumianowego.

(Dotyczy tylko ujemnego rozkładu dwumianowego) Możesz użyć wartości domyślnej -1 lub określić inną wartość parametru.

Ustawienia funkcji łączenia. Te ustawienia dotyczą funkcji łączenia, która wiąże zmienną zależną ze zmiennymi predyktora.

- **Funkcja łączenia.** Funkcja, która może być używana, jest jedną z następujących funkcji: **Identity**, **Inverse**, **Invnegative**, **Invsquare**, **Sqrt**, **Power**, **Oddspower**, **Log**, **Clog**, **Loglog**, **Cloglog**, **Logit** (domyślna), **Probit**, **Gaussit**, **Cauchit**, **Canbinom**, **Cangeom**, **Cannegbinom**.
- **Parametry.** (Dotyczy tylko funkcji łączenia Power lub Oddspower) Wartość parametru można określić, pod warunkiem że funkcją łączenia jest **Power** lub **Oddspower**. Określ wartość lub użyj wartości domyślnej 1.

Opcje zmiennej modelowania z użyciem uogólnionego modelu liniowego Netezza — interakcja

Panel Interakcja zawiera opcje określania interakcji (czyli efektu multiplikatywnego między zmiennymi wejściowymi).

Interakcja kolumny. Zaznacz to pole wyboru, aby określić interakcje między zmiennymi wejściowymi. Jeśli nie ma żadnych interakcji, pozostaw to pole puste.

Wprowadź interakcje do modelu, zaznaczając jedną zmienną lub większą liczbę zmiennych na liście źródłowej i przeciągając do listy interakcji. Typ tworzonej interakcji jest zależny od tego, do którego obszaru aktywnego zostanie przeciągnięte zaznaczenie.

- **Główne.** Zmienne odrzucane są wyświetlane jako osobne interakcje główne u dołu listy interakcji.
- **2. rzędu.** Wszystkie możliwe pary zmiennych odrzucanych pojawiają się jako interakcje 2. rzędu u dołu listy interakcji.
- **3. rzędu.** Wszystkie możliwe trójki zmiennych odrzucanych pojawiają się jako interakcje 3. rzędu u dołu listy interakcji.
- *****. Kombinacja wszystkich zmiennych odrzucanych pojawia się jako pojedyncza interakcja u dołu listy interakcji.

Uwzględnij wyraz wolny. Wyraz wolny jest zwykle uwzględniany w modelu. Jeśli można założyć, że dane przechodzą przez początek układu współrzędnych, wyraz wolny można wyłączyć z modelu.

Przyciski w oknie dialogowym

Przyciski po prawej stronie obszaru wyświetlania umożliwiają wprowadzanie zmian do składników używanych w modelu.



Rysunek 5. Przycisk Usuń

Umożliwia usuwanie składników z modelu poprzez wybranie składników przeznaczonych do usunięcia i kliknięcie przycisku usuwania.



Rysunek 6. Przycisk Reorganizacja

Umożliwia reorganizację składników w modelu poprzez wybranie składników przeznaczonych do reorganizacji i kliknięcie strzałki w górę lub w dół.



Rysunek 7. Przycisk interakcji niestandardowej

Dodaj składnik niestandardowy

Interakcje niestandardowe można określać w postaci $n1*x1*x1*x1*x1...$. Wybierz zmienną z listy **Zmienne**, kliknij przycisk ze strzałką w prawo, aby dodać zmienną do **Składnika niestandardowego**, kliknij opcję **Wg***, wybierz następną zmienną, kliknij przycisk ze strzałką w prawo i tak dalej. Po zbudowaniu interakcji niestandardowej kliknij opcję **Dodaj składnik**, aby zwrócić go do panelu interakcji.

Opcje uogólnionego modelu liniowego Netezza — opcje oceniania

Udostępnij do oceniania. W tym miejscu można ustawić wartości domyślne dla opcji oceniania, które będą widoczne w oknie dialogowym dla modelu użytkowego. Więcej informacji można znaleźć w temacie “Użytkowy uogólniony model liniowy Netezza — karta Ustawienia” na stronie 110.

- **Uwzględnij zmienne wejściowe.** Zaznacz to pole wyboru, jeśli zamierzasz wyświetlać zmienne wejściowe w wynikach modelu, a także w predykcjach.

Drzewa decyzyjne Netezza

Drzewo decyzyjne jest strukturą hierarchiczną, która reprezentuje model klasyfikacji. Za pomocą modelu drzewa decyzyjnego można opracować system klasyfikacji, aby przewidywać lub klasyfikować przyszłe obserwacje z zestawu danych uczących. Klasyfikacja przyjmuje postawę struktury drzewa, w której gałęzie reprezentują punkty podziału w klasyfikacji. Podziały rekurencyjnie rozdzielają dane na podgrupy, aż do osiągnięcia punktu zatrzymania. Węzły drzewa w punktach zatrzymania są znane pod nazwą **liście**. Każdy liść przypisuje etykietę znaną jako **etykieta klasy** do członków jego podgrupy lub klasy.

Wagi instancji i wagi klas

Domyślnie obowiązuje założenie, że wszystkie rekordy i klasy wejściowe mają równą wagę względną. Można to zmienić, przypisując niezależne wagi elementom każdej z tych grup lub obu tych grup. Taki sposób postępowania może być użyteczny, na przykład jeśli punkty danych w danych uczących nie są rozłożone realistycznie między kategorie. Wagi umożliwiają obciążenie modelu, dzięki czemu możliwe jest skompensowanie tych kategorii, które są gorzej reprezentowane w danych. Zwiększenie wagi wartości przewidywanej powinno spowodować wzrost procentu poprawnych predykcji dla tej kategorii.

W węzle modelowania Drzewo decyzyjne można określić dwa typy wag. **Wagi instancji** przypisują wagę dla każdego wiersza danych wejściowych. Wagi są zwykle określane jako 1.0 dla większości obserwacji, przy czym wyższe lub niższe wartości są nadawane tylko tym obserwacjom, które są bardziej lub mniej ważne niż większość, co przedstawia poniższa tabela.

Tabela 9. Przykład wagi instancji

ID rekordu	Zmienna przewidywana	Waga instancji
1	lekarstwoA	1.1
2	lekarstwoB	1.0
3	lekarstwoA	1.0
4	lekarstwoB	0.3

Wagi klas przypisują wagę do każdej kategorii zmiennej przewidywanej, co przedstawia poniższa tabela.

Tabela 10. Przykład wagi klasy

Klasa	Waga klasy
lekarstwoA	1.0
lekarstwoB	1.5

Oba typy wag mogą być używane jednocześnie, ale w takim przypadku są mnożone razem i używane jako wagi instancji. Dlatego, gdyby dwa poprzednie przykłady zostały użyte razem, algorytm użyłby wag instancji w sposób przedstawiony w poniższej tabeli.

Tabela 11. Przykład obliczenia wagi instancji

ID rekordu	Obliczenie	Waga instancji
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

Opcje zmiennej drzewa decyzyjnego Netezza

Na karcie Zmienne można zdecydować, czy mają zostać użyte ustawienia roli zmiennej już zdefiniowane w węzłach poprzedzających, czy też przypisania zmiennych mają zostać dokonane ręcznie.

Użyj wstępnie zdefiniowanych ról Ta opcja korzysta z ustawień roli (zmienne przewidywane, predyktory itd.) z poprzedzającego węzła Typ (lub z karty Typy poprzedzającego węzła źródłowego).

Użyj niestandardowych przypisań Tę opcję należy wybrać, aby ręcznie przypisać zmienne przewidywane, predyktory i inne role.

Zmienne Aby ręcznie przypisać pozycje z tej listy do różnych zmiennych ról po prawej stronie ekranu, należy użyć klawiszy strzałek. Ikony wskazują prawidłowe poziomy pomiaru dla każdego pola roli.

Aby wybrać wszystkie zmienne z listy, należy kliknąć przycisk **Wszystkie** lub kliknąć przycisk poziomu pojedynczego pomiaru, aby wybrać wszystkie zmienne dla tego poziomu pomiaru.

Przewidywana Należy wybrać jedną zmienną jako przewidywaną dla predykcji.

Identyfikator rekordu. Zmienna, która będzie używana jako unikatowy identyfikator rekordu. Wartości tej zmiennej muszą być unikalne dla każdego rekordu, np. muszą to być identyfikatory klientów.

Waga instancji. Określenie zmiennej w tym miejscu umożliwia użycie wag instancji (jedna waga na jeden wiersz danych wejściowych) zamiast — lub dodatkowo — domyślnych wag klas (jedna waga na kategorię dla zmiennej przewidywanej). Zmienna, którą określa użytkownik w tym miejscu, musi zawierać wagę liczbową dla każdego wiersza danych wejściowych. Więcej informacji można znaleźć w temacie “Wagi instancji i wagi klas” na stronie 87.

Predyktory (dane wejściowe). Umożliwia wybór zmiennej wejściowej lub kilku zmiennych. Działanie jest podobne jak w przypadku ustawienia roli zmiennej na *Zmienna wejściowa* w węzle Typ.

Opcje budowania drzewa decyzyjnego Netezza

Dla wzrostu drzewa dostępne są następujące opcje budowania:

Miara wzrostu. Te opcje kontrolują sposób wzrostu drzewa.

- **Miary zanieczyszczenia.** Ta miara ocenia najlepsze miejsce podziału drzewa. Jest to pomiar zmienności w podgrupie lub segmencie danych. Niski pomiar zanieczyszczenia wskazuje grupę, w której większość elementów ma podobne wartości dla kryterium lub zmiennej przewidywanej.

Obsługiwane pomiary są następujące: **Entropia** i **Gini**. Te miary bazują na prawdopodobieństwach członkostwa w kategorii dla gałęzi.

- **Maksymalna głębokość drzewa.** Maksymalna liczba liści, do której może rosnąć drzewo poniżej węzła głównego, czyli jest to liczba rekurencyjnych podziałów próby. Wartością domyślną tej właściwości jest 10, a wartość maksymalna, jaką można ustawić dla tej właściwości, wynosi 62.

Uwaga: Jeśli przeglądarka w modelu użytkowym przedstawia tekstową reprezentację modelu, wówczas wyświetlanych jest maksymalnie 12 poziomów drzewa.

Kryteria podziału. Te opcje kontrolują miejsce, w którym następuje zatrzymanie podziału drzewa.

- **Minimalne ulepszenie dla podziałów.** Minimalna wartość, o jaką musi zostać zmniejszone zanieczyszczenie, zanim w drzewie zostanie utworzony nowy podział. Celem budowy drzewa jest utworzenie podgrup o podobnych wartościach wyjściowych — czyli minimalizowanie zanieczyszczeń w ramach każdego węzła. Jeśli najlepszy podział dla gałęzi zmniejsza zanieczyszczenia o mniej, niż określono w kryteriach podziału, wówczas gałąź nie jest dzielona.
- **Minimalna liczba instancji dla podziału.** Minimalna liczba rekordów, jaka może zostać podzielona. Jeśli liczba pozostałych niepodzielonych rekordów jest mniejsza niż ta liczba, wówczas dalsze podziały nie są wykonywane. Tego pola można użyć, aby zapobiec tworzeniu małych podgrup w drzewie.

Statystyki. Ten parametr definiuje liczbę statystyk uwzględnionych w modelu. Wybierz jedną z poniższych opcji:

- **Wszystko.** Uwzględnione są wszystkie statystyki dotyczące kolumny i wszystkie statystyki dotyczące wartości.

Uwaga: Ten parametr zawiera maksymalną liczbę statystyk i może wpłynąć na wydajność systemu. Jeśli wyświetlanie modelu w formacie graficznym nie jest wymagane, należy określić wartość **Brak**.

- **Kolumny.** Uwzględnione są statystyki dotyczące kolumny.
- **Brak.** Uwzględnione są tylko statystyki, które są wymagane do oceny modelu.

Węzeł drzewa decyzyjnego Netezza — wagi klas

W tym miejscu można przypisywać wagi do poszczególnych klas. Domyślnie wartość 1 jest przypisywana do wszystkich klas, co powoduje, że uzyskują równe wagi. Określenie różnych wag liczbowych dla różnych etykiet klas oznacza, że algorytm otrzymuje instrukcję odpowiedniego obciążenia zestawów uczących konkretnych klas.

W celu zmiany wagi kliknij ją dwukrotnie w kolumnie **Waga** i wprowadź żądane zmiany.

Wartość. Zestaw etykiet klas uzyskany na podstawie możliwych wartości zmiennej przewidywanej.

Waga. Waga, jaka zostanie przypisana do konkretnej klasy. Przypisanie wyższej wagi do klasy powoduje, że model staje się bardziej czuły dla tej klasy w porównaniu do innych klas.

Wagi klas można używać w połączeniu z wagami instancji. Więcej informacji można znaleźć w temacie “Wagi instancji i wagi klas” na stronie 87.

Węzeł drzewa decyzyjnego Netezza — przycinanie drzewa

Za pomocą opcji przycinania można określić kryteria przycinania dla drzewa decyzyjnego. Celem przycinania jest zmniejszenie ryzyka przeuczenia poprzez usunięcie przerośniętych podgrup, które nie poprawiają oczekiwanej dokładności w przypadku nowych danych.

Miara przycinania. Domyślna miara przycinania — **Dokładność** — zapewnia, że oszacowana dokładność modelu mieści się w dozwolonych limitach po usunięciu liścia z drzewa. Jeśli podczas przycinania chcesz uwzględnić wagi klas, użyj alternatywnej opcji **Ważona dokładność**.

Dane dla przycinania. W celu oszacowania oczekiwanej dokładności nowych danych można użyć części lub wszystkich danych uczących. Alternatywnie do tego celu można użyć osobnego zestawu danych do przycinania z podanej tabeli.

- **Użyj wszystkich danych uczących.** Ta opcja (domyślna) powoduje, że wykorzystywane są wszystkie dane uczące w celu oszacowania dokładności modelu.
- **Użyj % danych uczących do przycinania.** Ta opcja umożliwia podział danych na dwa zestawy — jeden jest używany do nauki, a drugi do przycinania, przy czym określony w opcji procent dotyczy danych do przycinania. Jeśli chcesz określić wartość początkową generatora liczb losowych, aby upewnić się, że dane zostaną podzielone tak samo za każdym razem, gdy uruchomisz strumień, wybierz opcję **Replikuj wyniki**. Możesz określić wartość całkowitą w polu **Wartość startowa generatora użyta do przycinania** lub kliknąć opcję **Utwórz**, co spowoduje utworzenie pseudolosowej wartości całkowitej.
- **Użyj danych z istniejącej tabeli.** Określ nazwę tabeli osobnego zestawu danych do przycinania w celu oszacowania dokładności modelu. Takie postępowanie jest traktowane jako bardziej niezawodne niż korzystanie z danych uczących. Jednak użycie tej opcji może spowodować usunięcie dużego zestawu danych z zestawu uczącego, co spowoduje obniżenie jakości drzewa decyzyjnego.

Netezza Linear Regression

Modele liniowe przewidują przewidywaną zmienną ilościową na podstawie liniowych relacji między przewidywaną a jednym predyktorem lub większą ich liczbą. Modele regresji liniowej są stosunkowo proste i zapewniają łatwy w interpretacji wzór matematyczny do oceniania, jednak są ograniczone do bezpośredniego modelowania zależności liniowych. Modele liniowe są szybkie, wydajne i łatwe w użyciu, ale zakres ich zastosowań jest ograniczony w porównaniu do modeli zwracanych przez bardziej wyrafinowane algorytmy regresji.

Opcje budowania regresji liniowej Netezza

Na karcie Opcje budowania można ustawić wszystkie opcje budowania modelu. Można oczywiście po prostu kliknąć przycisk **Uruchom** w celu zbudowania modelu ze wszystkimi opcjami domyślnymi, lecz w normalnej sytuacji użytkownik potrzebuje zwykle dostosować proces budowy do swoich celów.

Używaj dekompozycji wartości osobliwych w celu rozwiązywania równań. Korzystanie z macierzy dekompozycji wartości osobliwych zamiast z macierzy oryginalnych ma tę zaletę, że jest bardziej odporne na błędy liczbowe i może przyspieszyć obliczenia.

Uwzględnij wyraz wolny w modelu. Uwzględnienie wyrazu wolnego zwiększa ogólną dokładność rozwiązania.

Oblicz diagnostyki modelu. Ta opcja powoduje obliczenie szeregu diagnostyk względem modelu. Wyniki są zapisywane w macierzach lub tabelach do przejrzania później. Diagnostyki obejmują następujące współczynniki: R-kwadrat, suma kwadratów reszt, oszacowanie wariancji, odchylenie standardowe, wartość p oraz wartość t .

Te diagnostyki dotyczą prawidłowości i użyteczności modelu. W celu upewnienia się, że dane spełniają założenia liniowości, należy uruchamiać osobne diagnostyki względem danych bazowych.

Netezza KNN

Analiza najbliższego sąsiedztwa jest metodą klasyfikacji obserwacji na podstawie ich podobieństwa do innych obserwacji. Zostało to opracowane w nauczaniu maszynowym jako sposób rozpoznawania wzorców danych bez konieczności zapewnienia dokładnej zgodności z jakimikolwiek zapamiętanymi wzorcami lub obserwacjami. Podobne obserwacje znajdują się blisko siebie, a niepodobne – daleko. Zatem odległość między dwoma obserwacjami stanowi miarę ich niepodobieństwa.

Obserwacje znajdujące się blisko siebie nazywają się „sąsiedztwem”. Podczas prezentacji nowej (wstrzymanej) obserwacji obliczana jest odległość od każdej obserwacji modelu. Zostaje określona klasyfikacja najbardziej podobnych obserwacji najbliższego sąsiedztwa, a nowa obserwacja zostaje umieszczona w kategorii, która zawiera największą liczbę obserwacji najbliższego sąsiedztwa.

Można określić liczbę najbliższych elementów sąsiednich do analizowania; ta wartość to k . Rysunki przedstawiają, jak nowa obserwacja będzie sklasyfikowana za pomocą dwóch różnych wartości k . Jeśli $k = 5$, nowa obserwacja jest umieszczana w kategorii 1 , ponieważ większość najbliższych elementów sąsiednich należy do kategorii 1 . Jeśli jednak $k = 9$, nowa obserwacja jest umieszczana w kategorii 0 , ponieważ większość najbliższych elementów sąsiednich należy do kategorii 0 .

Analiza najbliższego sąsiedztwa może być również użyta do obliczania docelowych wartości ilościowych. W tej sytuacji do uzyskania przewidywanej wartości dla nowej obserwacji stosowana jest docelowa wartość średniej lub mediany najbliższych sąsiadów.

Opcje modelu KNN Netezza — informacje ogólne

Na karcie Opcje modelu — Opcje ogólne można zdecydować o wyborze nazwy dla modelu lub o jej wygenerowaniu automatycznie. Można także ustawić opcje kontrolujące sposób obliczania najbliższych sąsiadów i ustawiać opcje poprawy wydajności i dokładności modelu.

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu, w przypadkach gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Sąsiedzi

Miara odległości. Metoda używana w celu pomiaru odległości między punktami danych; większa odległość oznacza większe niepodobieństwo. Opcje są następujące:

- **Euklidesowa.** (opcja domyślna) Odległość między dwoma punktami wyliczona jako długość łączącej je linii prostej.
- **Manhattan.** Odległość między dwoma punktami jest sumą bezwzględnych różnic między ich współrzędnymi.
- **Canberra.** Podobna do odległości Manhattan, ale bardziej czuła w przypadku punktów danych położonych bliżej początku układu współrzędnych.
- **Maksimum.** Odległość między dwoma punktami jest obliczana jako największa z różnic wzdłuż dowolnej współrzędnej.

Liczba najbliższych sąsiadów (k). Liczba najbliższych sąsiadów dla konkretnej obserwacji. Należy pamiętać, że większa liczba obserwacji najbliższego sąsiedztwa nie zawsze oznacza dokładniejszy model.

Wybranie opcji *k* zapewnia równowagę między zapobieganiem przeuczeniu (to może być szczególnie istotne w przypadku zaszumionych danych) i rozdzielaniem (uzyskiwaniem różnych predykcji w przypadku podobnych instancji). Zwykle konieczne jest dostosowanie wartości *k* dla każdego zestawu danych, przy czym typowe wartości należą do zakresu od 1 do kilku tuzinów.

Zwiększanie wydajności i dokładności

Standaryzacja pomiarów przed obliczeniem. Wybranie tej opcji powoduje standaryzację pomiarów dla ciągłych danych wejściowych przed obliczeniem wartości odległości.

Użyj zestawów podstawowych w celu poprawy wydajności w przypadku dużych zestawów danych. Wybranie tej opcji powoduje używanie próbkowania zestawu podstawowego w celu przyspieszenia obliczeń, gdy używane są duże zestawy danych.

Opcje modelu KNN Netezza — opcje oceny

Na karcie Opcje modelu — opcje oceny można ustawić wartość domyślną dla opcji oceny i przypisać wartości względne do pojedynczych klas.

Uczyń dostępnym dla scoringu

Uwzględnij zmienne wejściowe. Określa, czy zmienne wejściowe są domyślnie uwzględnione w ocenie.

Wagi klas

Tej opcji należy użyć, jeśli wymagana jest zmiana względnej wartości poszczególnych klas podczas budowania modelu.

Uwaga: ta opcja jest aktywna tylko wówczas, gdy do klasyfikacji używany jest algorytm KNN. Jeśli wykonywana jest regresja (czyli zmienna przewidywana jest typu ciągłego), ta opcja jest wyłączona.

Domyślnie wartość 1 jest przypisywana do wszystkich klas, co powoduje, że uzyskują równe wagi. Określenie różnych wag liczbowych dla różnych etykiet klas oznacza, że algorytm otrzymuje instrukcję odpowiedniego obciążenia zestawów uczących konkretnych klas.

W celu zmiany wagi kliknij ją dwukrotnie w kolumnie **Waga** i wprowadź żądane zmiany.

Wartość. Zestaw etykiet klas uzyskany na podstawie możliwych wartości zmiennej przewidywanej.

Waga. Waga, jaka zostanie przypisana do konkretnej klasy. Przypisanie wyższej wagi do klasy powoduje, że model staje się bardziej czuły dla tej klasy w porównaniu do innych klas.

Netezza K-Means

Węzeł K-średnie implementuje algorytm *k*-średnie, który zapewnia metodę analizy skupień. Za pomocą tego węzła można grupować zestaw danych do odrębnych grup.

Ten algorytm jest algorytmem grupowania zależnym od odległości, którego działanie opiera się na metryce (funkcji) odległości w celu pomiaru podobieństwa między punktami danych. Punkty danych są przypisane do najbliższego skupienia zgodnie z używaną metryką odległości.

Algorytm działa, wykonując kilka iteracji tego samego podstawowego procesu, w którym każda instancja ucząca jest przypisywana do najbliższego skupienia (względem określonej funkcji odległości, stosowanej do instancji i centrum skupienia). Wszystkie centra skupień są następnie obliczane ponownie jako wektory wartości atrybutu średniej instancji przypisanych do poszczególnych skupień.

Opcje zmiennej K-średnie Netezza

Na karcie Zmienne można zdecydować, czy mają zostać użyte ustawienia roli zmiennej już zdefiniowane w węzłach poprzedzających, czy też przypisania zmiennych mają zostać dokonane ręcznie.

Użyj wstępnie zdefiniowanych ról. Ta opcja korzysta z ustawień roli (zmienne przewidywane, predyktory itd.) z poprzedzającego węzła Typ (lub z karty Typy poprzedzającego węzła źródłowego).

Użyj niestandardowych przypisań. Opcję tę należy wybrać w celu ręcznego przypisania zmiennych przewidywanych, predyktorów oraz innych ról na tym ekranie.

Zmienne. Aby ręcznie przypisać pozycje z tej listy do różnych zmiennych ról po prawej stronie ekranu, należy użyć klawiszy strzałek. Ikony wskazują prawidłowe poziomy pomiaru dla każdego pola roli.

Kliknij przycisk **Wszystkie**, aby wybrać wszystkie pola na liście, lub kliknij przycisk poziomu pojedynczego pomiaru, aby wybrać wszystkie pola, które posiadają ten poziom pomiaru.

Identyfikator rekordu. Zmienna, która będzie używana jako unikatowy identyfikator rekordu.

Predyktory (dane wejściowe). Wybierz jedną lub więcej zmiennych jako dane wejściowe dla predykcji.

Karta Opcje budowania K-średnie Netezza

Ustawienie opcji budowania umożliwia dostosowanie modelu do konkretnego celu.

Jeśli zamierzasz zbudować model z opcjami domyślnymi, kliknij opcję **Wykonaj**.

Miara odległości. Ten parametr definiuje metodę pomiaru odległości między punktami danych. Większe odległości oznaczają większe niepodobieństwa. Wybierz jedną z poniższych opcji:

- **Euklidesowa.** Odległość euklidesowa jest odległością „w linii prostej” pomiędzy dwoma punktami danych.
- **Znormalizowana euklidesowa.** Miara Znormalizowana euklidesowa jest podobna do miary euklidesowej, ale jest znormalizowana przez kwadrat odchylenia standardowego. Miara Znormalizowana euklidesowa, w przeciwieństwie do miary euklidesowej, jest również skalo-niezmiennicza.
- **Mahalanobisa.** Miara Mahalanobisa jest uogólnioną miarą euklidesową, która uwzględnia korelacje danych wejściowych. Miara Mahalanobisa, podobnie jak miara Znormalizowana euklidesowa, jest skalo-niezmiennicza.
- **Manhattan.** Miara Manhattan jest odległością między dwoma punktami danych, która jest obliczana jako suma bezwzględnej różnicy ich współrzędnych.

- **Canberra.** Miara Canberra jest podobna do miary Manhattan, ale bardziej czuła w przypadku punktów danych położonych bliżej początku układu współrzędnych.
- **Maksimum.** Odległość Maksimum to odległość między dwoma punktami, która jest obliczana jako największa z różnic wzdłuż dowolnej współrzędnej.

Liczba grup. Ten parametr definiuje liczbę skupień do utworzenia.

Maksymalna liczba iteracji. Algorytm wykonuje kilka iteracji tego samego procesu. Ten parametr definiuje liczbę iteracji, po których szkolenie modelu jest zatrzymywane.

Statystyki. Ten parametr definiuje liczbę statystyk uwzględnionych w modelu. Wybierz jedną z poniższych opcji:

- **Wszystko.** Uwzględnione są wszystkie statystyki dotyczące kolumny i wszystkie statystyki dotyczące wartości.

Uwaga: Ten parametr zawiera maksymalną liczbę statystyk i może wpłynąć na wydajność systemu. Jeśli wyświetlanie modelu w formacie graficznym nie jest wymagane, należy określić wartość **Brak**.

- **Kolumny.** Uwzględnione są statystyki dotyczące kolumny.
- **Brak.** Uwzględnione są tylko statystyki, które są wymagane do oceny modelu.

Replikacja wyników. Zaznacz to pole wyboru, jeśli zamierzasz ustawić wartość początkową generatora liczb losowych, aby replikować wyniki. Możesz określić wartość całkowitą lub utworzyć pseudolosową wartość całkowitą, klikając opcję **Utwórz**.

Netezza Naive Bayes

Naive Bayes to dobrze znany algorytm stosowany w przypadku problemów z klasyfikacją. Model jest określany jako *naiwny*, ponieważ traktuje wszystkie proponowane zmienne predykcji jako niezależne od siebie. Naive Bayes to szybki, skalowalny algorytm, który oblicza prawdopodobieństwa warunkowe kombinacji atrybutów i atrybutu przewidywanego. Na podstawie danych uczących wyznaczane jest niezależne prawdopodobieństwo. To prawdopodobieństwo określa wiarygodność każdej klasy przewidywanej z uwzględnieniem wystąpienia każdej kategorii wartości z poszczególnych zmiennych wejściowych.

Netezza Bayes Net

Sieć bayesowska jest modelem prezentującym zmienne w zbiorze danych, a także probabilistyczne i warunkowe współzależności między tymi zmiennymi. Za pomocą węzła Netezza Bayes Net możesz utworzyć model prawdopodobieństwa przez połączenie zaobserwowanych i zarejestrowanych dowodów ze „zdroworozsądkową” wiedzą rzeczywistą w celu ustanowienia prawdopodobieństwa występowania zdarzeń na podstawie pozornie niepowiązanych ze sobą atrybutów.

Opcje zmiennej sieci Bayesa Netezza

Na karcie Zmienne można zdecydować, czy mają zostać użyte ustawienia roli zmiennej już zdefiniowane w węzłach poprzedzających, czy też przypisania zmiennych mają zostać dokonane ręcznie.

W przypadku tego węzła zmienna przewidywana jest wymagana tylko do oceny, dlatego nie jest wyświetlana na tej karcie. Zmienną przewidywaną można ustawić lub zmienić w węźle Typ, na karcie opcji modelu tego węzła, a także na karcie Ustawienia modelu użytkowanego. Więcej informacji można znaleźć w temacie “Model użytkowy Sieci Bayesa Netezza — karta Ustawienia” na stronie 105.

Użyj wstępnie zdefiniowanych ról. Ta opcja korzysta z ustawień roli (zmienne przewidywane, predyktory itd.) z poprzedzającego węzła Typ (lub z karty Typy poprzedzającego węzła źródłowego).

Użyj niestandardowych przypisań. Opcję tę należy wybrać w celu ręcznego przypisania zmiennych przewidywanych, predyktorów oraz innych ról na tym ekranie.

Zmienne. Aby ręcznie przypisać pozycje z tej listy do różnych zmiennych ról po prawej stronie ekranu, należy użyć klawiszy strzałek. Ikony wskazują prawidłowe poziomy pomiaru dla każdego pola roli.

Kliknij przycisk **Wszystkie**, aby wybrać wszystkie pola na liście, lub kliknij przycisk poziomu pojedynczego pomiaru, aby wybrać wszystkie pola, które posiadają ten poziom pomiaru.

Predyktory (dane wejściowe). Wybierz jedną lub więcej zmiennych jako dane wejściowe dla predykcji.

Opcje budowania zmiennej Sieci Bayesa Netezza

Na karcie Opcje budowania można ustawić wszystkie opcje budowania modelu. Można oczywiście po prostu kliknąć przycisk **Uruchom** w celu zbudowania modelu ze wszystkimi opcjami domyślnymi, lecz w normalnej sytuacji użytkownik potrzebuje zwykle dostosować proces budowy do swoich celów.

Indeks podstawowy. Identyfikator liczbowy, który zostanie przypisany do pierwszego atrybutu (zmiennej wejściowej) w celu łatwiejszego zarządzania wewnętrznego.

Wielkość próby. Wielkość próby przyjmowana, gdy liczba atrybutów jest tak duża, że spowodowałaby niedopuszczalnie długi czas przetwarzania.

Wyświetlaj dodatkowe informacje podczas wykonania. Jeśli to pole wyboru jest zaznaczone (domyślnie), wówczas dodatkowe informacje o postępie są wyświetlane w oknie dialogowym komunikatu.

Netezza Time Series

Szereg czasowy to sekwencja wartości liczbowych, które są mierzone w kolejnych (choć niekoniecznie regularnych) punktach czasu — na przykład codzienne ceny akcji na giełdzie lub dane dotyczące sprzedaży tygodniowej. Analizowanie takich danych może być użyteczne, na przykład w celu wyróżnienia zjawisk, takich jak trendy i sezonowość (powtarzający się wzorzec), a także przewidywania przyszłych zachowań na podstawie zdarzeń z przeszłości.

Szereg czasowy Netezza obsługuje następujące algorytmy dotyczące szeregu czasowego.

- analiza spektralna
- wygładzanie wykładnicze
- Autoregresyjna zintegrowana średnia ruchoma (ARIMA)
- dekompozycja trendu sezonowego

Te algorytmy rozbijają szereg czasowy na trend i składnik sezonowy. Te składniki są następnie analizowane w celu zbudowania modelu, który może być używany w celu przewidywania.

Analiza spektralna jest używana do identyfikacji zachowań okresowych w szeregu czasowym. Gdy szereg czasowy jest złożony z wielu bazowych okresowości lub gdy dane zawierają znaczącą ilość szumu losowego, wówczas analiza spektralna stanowi najłatwiejszy sposób identyfikacji składników okresowych. Ta metoda wykrywa częstotliwości zachowania okresowego, przeprowadzając transformację szeregu z domeny czasu do szeregu w domenie częstotliwości.

Wygładzanie wykładnicze to metoda prognozowania wykorzystująca wartości ważone poprzednich obserwacji szeregu do predykcji przyszłych wartości. Wygładzanie wykładnicze sprawia, że wpływ obserwacji zmniejsza się w czasie w sposób wykładniczy. Ta metoda prognozuje jeden punkt naraz, dopasowując jego prognozy w miarę dochodzenia nowych danych, uwzględniając dodatki, trend i sezonowość.

Modele **ARIMA** oferują bardziej wyrafinowane metody modelowania składników trendu i sezonowości niż modele wygładzania wykładniczego. Ta metoda obejmuje jawne określanie kolejności autoregresji i średnich ruchomych, a także stopnia różnicowania.

Uwaga: W praktyce modele ARIMA są najbardziej użyteczne w sytuacji, gdy pożądane jest uwzględnienie predyktorów, które mogą pomóc w wyjaśnieniu zachowania prognozowanego szeregu, takich jak liczba rozesłanych katalogów czy liczba odwiedzin na stronie WWW firmy. Modele wygładzania wykładniczego opisują przebieg szeregu czasowego bez próby zrozumienia przyczyn takiego przebiegu.

Dekompozycja trendu sezonowego usuwa zachowanie okresowe z szeregu czasowego w celu przeprowadzenia analizy trendu, a następnie wybiera podstawowy kształt dla trendu, np. funkcję kwadratową. Te kształty podstawowe mają szereg parametrów, których wartości są ustalane w celu zminimalizowania średniego błędu kwadratowego reszt (czyli różnic między wartościami dopasowanymi a obserwowanymi w szeregu czasowym).

Interpolacja wartości w szeregu czasowym Netezza

Interpolacja to proces szacowania i wstawiania wartości brakujących do danych szeregu czasowego.

Jeśli przedziały szeregu czasowego są regularne, ale niektóre wartości są nieobecne, wówczas wartości brakujące mogą zostać oszacowane przy użyciu interpolacji liniowej. Rozważmy następujący szereg przylotów pasażerów na terminal lotniskowy (z okresu miesiąca).

Tabela 12. Miesięczne przyloty na terminal pasażerski

Miesiąc	Pasażerowie
3	3 500 000
4	3 900 000
5	-
6	3 400 000
7	4 500 000
8	3 900 000
9	5 800 000
10	6 000 000

W tym przypadku interpolacja liniowa oszacowałaby, że brakująca wartość dla miesiąca 5 wynosi 3 650 000 (punkt środkowy między miesiącami 4 i 6).

Przedziały nieregularne są obsługiwane w inny sposób. Rozważmy następujący szereg odczytów temperatury.

Tabela 13. Odczyty temperatury

Data	Czas	Temperatura
2011-07-24	7:00	57
2011-07-24	14:00	75
2011-07-24	21:00	72
2011-07-25	7:15	59
2011-07-25	14:00	77
2011-07-25	20:55	74
2011-07-27	7:00	60
2011-07-27	14:00	78
2011-07-27	22:00	74

Mamy odczyty wykonywane w trzech punktach w ciągu dnia, ale o różnych godzinach. Tylko część z nich jest taka sama między poszczególnymi dniami. Ponadto tylko dwa dni są ciągłe.

Z taką sytuacją można sobie poradzić na dwa sposoby: poprzez obliczanie agregacji lub określenie rozmiaru kroku.

Agregacje mogą być agregacjami dziennymi obliczanymi zgodnie z wzorem opartym na semantycznej znajomości danych. Zastosowanie takiego rozwiązania mogłoby zwrócić następujący zestaw danych.

Tabela 14. Odczyty temperatury (zagregowane)

Data	Czas	Temperatura
2011-07-24	24:00	69
2011-07-25	24:00	71
2011-07-26	24:00	puste
2011-07-27	24:00	72

Alternatywnie algorytm może traktować szereg jako odrębny szereg i określić odpowiedni rozmiar kroku. W takim przypadku rozmiar kroku określony przez algorytm może wynosić 8 godzin, co spowoduje uzyskanie poniższych danych.

Tabela 15. Odczyty temperatury z obliczonym rozmiarem kroku

Data	Czas	Temperatura
2011-07-24	6:00	
2011-07-24	14:00	75
2011-07-24	22:00	
2011-07-25	6:00	
2011-07-25	14:00	77
2011-07-25	22:00	
2011-07-26	6:00	
2011-07-26	14:00	
2011-07-26	22:00	
2011-07-27	6:00	
2011-07-27	14:00	78
2011-07-27	22:00	74

W tym przypadku cztery odczyty są zgodne z pomiarami oryginalnymi, ale za pomocą innych znanych wartości z szeregu oryginalnego można ponownie obliczyć wartości brakujące, stosując interpolacje.

Opcje zmiennych szeregu czasowego Netezza

Na karcie Zmienne określ role dla zmiennych wejściowych w danych źródłowych.

Zmienne. Aby ręcznie przypisać pozycje z tej listy do różnych zmiennych ról po prawej stronie ekranu, należy użyć klawiszy strzałek. Ikony wskazują prawidłowe poziomy pomiaru dla każdego pola roli.

Zmienna przewidywana. Wybierz jedną zmienną jako zmienną przewidywaną dla predykcji. Musi to być zmienna, w której poziom pomiaru jest ciągły.

(Predyktor) Punkty czasu. (wymagana) Zmienna wejściowa zawierająca wartości daty lub godziny dla szeregu czasowego. Musi to być zmienna o poziomie pomiaru ciągłym lub jakościowym, a dane muszą być zapisywane jako typ Data, Godzina, Znacznik czasu lub Liczbowe. Typ zapisu danych zmiennej określony w tym miejscu definiuje typ wejściowy dla niektórych zmiennych na innych kartach tego węzła modelowania.

(Predyktor) Identyfikator szeregu czasowego (Wg). Zmienna zawierająca identyfikatory szeregu czasowego; należy jej użyć, jeśli dane wejściowe zawierają więcej niż jeden szereg czasowy.

Opcje budowania szeregu czasowego Netezza

Dostępne są dwa poziomy opcji budowania:

- Podstawowy — ustawienia dla wyboru algorytmu, interpolacji i używanego zakresu czasu.
- Zaawansowany — ustawienia prognozowania

W niniejszej sekcji opisano opcje podstawowe.

Na karcie Opcje budowania można ustawić wszystkie opcje budowania modelu. Można oczywiście po prostu kliknąć przycisk **Uruchom** w celu zbudowania modelu ze wszystkimi opcjami domyślnymi, lecz w normalnej sytuacji użytkownik potrzebuje zwykle dostosować proces budowy do swoich celów.

Algorytm

Są to ustawienia dotyczące algorytmu szeregu czasowego, który będzie używany.

Nazwa algorytmu. Wybierz algorytm szeregu czasowego, którego chcesz użyć. Dostępne są algorytmy **Analiza spektralna**, **Wygladzanie wykładnicze** (domyślny), **ARIMA** oraz **Dekompozycja trendu sezonowego**. Więcej informacji można znaleźć w temacie “Netezza Time Series” na stronie 94.

Trend. (Dotyczy tylko wygladzania wykładniczego) Proste wygladzanie wykładnicze nie działa dobrze, jeśli w szeregu czasowym ujawnia się trend. Za pomocą tego pola można określić trend (jeśli istnieje), dzięki czemu algorytm będzie mógł go uwzględnić.

- **Określona przez system.** (Opcja domyślna) System podejmuje próbę znalezienia wartości optymalnej dla tego parametru.
- **Brak(N).** W szeregu czasowym nie ujawnia się trend.
- **Addytywny(A).** Trend, który stopniowo wzrasta w czasie.
- **Wygasający addytywny(DA).** Trend addytywny, który stopniowo zanika.
- **Multiplikatywny(M).** Trend, który narasta w czasie, zwykle szybciej niż trend stabilnie addytywny.
- **Wygasający multiplikatywny(DM).** Trend multiplikatywny, który stopniowo zanika.

Sezonowość. (Dotyczy tylko wygladzania wykładniczego) W tym polu można określić, czy w szeregu czasowym ujawniają się wzorce sezonowe w danych.

- **Określona przez system.** (Opcja domyślna) System podejmuje próbę znalezienia wartości optymalnej dla tego parametru.
- **Brak(N).** W szeregu czasowym nie ujawniają się wzorce sezonowe.
- **Addytywny(A).** Wzorec fluktuacji sezonowych wykazuje trend stabilnie wzrastający w czasie.
- **Multiplikatywny(M).** Tak samo jak sezonowy addytywny, ale dodatkowo amplituda (odległość między wysokimi i niskimi punktami) fluktuacji sezonowych zwiększa się w porównaniu do ogólnego trendu w górę fluktuacji.

Dla ARIMA użyj ustawień określonych przez system. (Dotyczy tylko ARIMA) Tę opcję należy wybrać, jeśli to system ma ustalić ustawienia dla algorytmu ARIMA.

Określ. (Dotyczy tylko ARIMA) W celu ręcznego określenia ustawień ARIMA należy wybrać tę opcję i kliknąć przycisk.

Interpolacja

Jeśli w danych źródłowych szeregu czasowego występują wartości brakujące, w celu wypełnienia przerw w danych wybierz metodę wstawiania wartości oszacowanych. Więcej informacji można znaleźć w temacie “Interpolacja wartości w szeregu czasowym Netezza” na stronie 95.

- **Liniowy.** Tę metodę należy wybrać, jeśli przedziały szeregu czasowego są regularne, ale niektóre wartości są nieobecne.
- **Wykładnicze krzywe sklejane.** Ta opcja dopasowuje krzywą wygładzoną, w której wartości znanego punktu danych zwiększają się lub zmniejszają w szybkim tempie.
- **Sześciennne krzywe sklejane.** Dopasowuje krzywą wygładzoną do znanych punktów danych w celu oszacowania wartości brakujących.

Zakres czasu

W tym miejscu można wybrać, czy w szeregu czasowym w celu utworzenia modelu używany będzie pełny zakres danych, czy ciągły podzbiór tych danych. Poprawne dane wejściowe dla tych pól są zdefiniowane przez typ zapisu danych zmiennej określonej dla punktów czasowych na karcie Zmienne. Więcej informacji można znaleźć w temacie “Opcje zmiennych szeregu czasowego Netezza” na stronie 96.

- **Użyj najwcześniejszych lub najpóźniejszych czasów dostępnych w danych.** Wybierz tę opcję, jeśli chcesz użyć pełnego zakresu danych szeregu czasowego.
- **Określ okno czasowe.** Wybierz tę opcję, jeśli chcesz użyć tylko części szeregu czasowego. W celu określenia granic użyj pól **Najwcześniejszy czas (od)** oraz **Najpóźniejszy czas (do)**.

Struktura ARIMA

Określ wartości różnych niesezonowych i sezonowych składników modelu ARIMA. W każdym przypadku ustaw operator na = (równość) lub <= (mniej niż lub równość), a następnie określ wartość w sąsiednim polu. Wartości muszą być nieujemnymi liczbami całkowitymi określającymi stopnie.

Niesezonowy. Wartości dla różnych niesezonowych składników modelu.

- **Stopnie autokorelacji (p).** Liczba rzędów autoregresji w modelu. Rzędy autoregresji określają, które z poprzednich wartości są używane do przewidywania bieżących wartości. Na przykład rząd autoregresji 2 oznacza, że do przewidywania bieżącej wartości zostanie użyta wartość dwu okresów czasu szeregu w przeszłości.
- **Wyrowadzenie (d).** Określa rząd różnicowania stosowany względem szeregu przed oszacowaniem modeli. Różnicowanie jest niezbędne w przypadku obecności trendów (szeregi z trendami są zwykle niestacjonarne, zaś modelowanie ARIMA zakłada stacjonarność) i służy do usuwania ich wpływu. Rząd różnicowania odpowiada stopniowi trendu szeregu — różnicowanie pierwszego rzędu jest uwzględniane w przypadku trendów liniowych, różnicowanie drugiego rzędu w przypadku trendów kwadratowych itd.
- **Średnia ruchoma (q).** Liczba rzędów średniej ruchomej w modelu. Rzędy średniej ruchomej określają, w jaki sposób odchylenia od średniej szeregu dla poprzednich wartości są używane do przewidywania bieżących wartości. Na przykład rzędy średniej ruchomej 1 i 2 oznaczają, że odchylenia od wartości średniej szeregu z każdego z dwu ostatnich okresów czasu będą uwzględniane podczas przewidywania bieżących wartości szeregu.

Sezonowe. Składniki autokorelacja sezonowa (SP), wyrowadzenie (SD) i średnia ruchoma (SQ) odgrywają te same role, jak ich niesezonowe odpowiedniki. W przypadku rzędów sezonowości na bieżące wartości szeregu wpływają jednak poprzednie wartości szeregu, rozdzielone jednym lub większą liczbą okresów sezonowości. Na przykład w przypadku danych miesięcznych (okres sezonowości 12) rząd sezonowości 1 oznacza, że na bieżącą wartość szeregu wpływa 12 okresów wartości szeregu poprzedzających okres bieżący. Rząd sezonowości 1, w przypadku danych miesięcznych jest taki sam, jak w przypadku rzędu niesezonowości wynoszącego 12.

Ustawienia sezonowe są uwzględniane tylko wtedy, jeśli zostanie wykryta sezonowość lub jeśli użytkownik określi ustawienia okresu na karcie Zaawansowane.

Opcje budowania szeregu czasowego Netezza — Zaawansowane

Ustawień zaawansowanych można używać w celu określania opcji prognozowania.

W przypadku opcji budowania modelu użyj ustawień określonych przez system. Tę opcję należy wybrać, jeśli to system ma ustalić ustawienia zaawansowane.

Określ. Tę opcję należy wybrać, jeśli opcje zaawansowane będą określane ręcznie. (Ta opcja jest niedostępna, jeśli algorytmem jest analiza spektralna).

- **Okres/Jednostki dla okresu.** Okres, po jakim pewne zachowanie charakterystyczne dla szeregu czasowego powtarza się. Na przykład w szeregu czasowym sprzedaży tygodniowej można określić 1 dla okresu i Tygodnie jako jednostki. **Okres** musi być nieujemną liczbą całkowitą; **Jednostki dla okresu** mogą być jedną z następujących: **Milisekundy, Sekundy, Minuty, Godziny, Dni, Tygodnie, Kwartały** lub **Lata**. Nie należy ustawiać **Jednostek dla okresu**, jeśli **Okres** nie jest ustawiony lub jeśli typ czasu nie jest liczbowy. Jeśli jednak **Okres** zostanie określony, należy określić również **Jednostki dla okresu**.

Ustawienia prognozy. Prognozy można wykonywać do konkretnego punktu czasowego albo w konkretnych punktach czasu. Poprawne dane wejściowe dla tych pól są zdefiniowane przez typ zapisu danych zmiennej określonej dla punktów czasowych na karcie Zmienne. Więcej informacji można znaleźć w temacie “Opcje zmiennych szeregu czasowego Netezza” na stronie 96.

- **Horyzont prognoz.** Wybierz tę opcję, jeśli zamierzasz określić tylko punkt końcowy prognozowania. Prognozy będą opracowywane tylko do tego punktu czasowego.
- **Czasy prognoz.** Wybierz tę opcję, aby określić co najmniej jeden punkt w czasie, w którym wykonywane będą prognozy. Kliknij przycisk **Dodaj**, aby dodać nowy wiersz do tabeli punktów czasowych. Aby usunąć wiersz, wybierz nowy wiersz i kliknij opcję **Usuń**.

Opcje modelu szeregów czasowych Netezza

Na karcie Opcje modelu można zdecydować o wyborze nazwy dla modelu lub o jej wygenerowaniu automatycznie. Dla opcji wyników modelu można również ustawić wartości domyślne.

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu, w przypadkach gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Udostępnij do oceniania. W tym miejscu można ustawić wartości domyślne dla opcji oceniania, które będą widoczne w oknie dialogowym dla modelu użytkowego.

- **W wyniku uwzględnij wartości historyczne.** Domyślnie wynik modelu nie obejmuje wartości z danych historycznych (tych, które są używane do przeprowadzenia predykcji). Zaznacz to pole wyboru, aby uwzględnić te wartości.
- **W wyniku uwzględnij wartości interpolowane.** Jeśli w wyniku zamierzasz uwzględnić wartości historyczne, zaznacz to pole, jeśli zamierzasz także uwzględnić wartości interpolowane (jeśli istnieją). Zwróć uwagę na to, że interpolacja działa tylko w przypadku danych historycznych, dlatego to pole jest niedostępne, jeśli nie wybrano opcji **W wyniku uwzględnij wartości historyczne**. Więcej informacji można znaleźć w temacie “Interpolacja wartości w szeregu czasowym Netezza” na stronie 95.

Netezza TwoStep

Węzeł Dwustopniowa implementuje algorytm dwustopniowy, który udostępnia metodę grupowania danych z dużych zestawów danych.

Za pomocą tego węzła można grupować dane, gdy uwzględniane są dostępne zasoby, np. ograniczona ilość pamięci i czasu.

Algorytm Dwustopniowa jest algorytmem eksploracji w bazie danych, który grupuje dane w następujący sposób:

1. Tworzone jest drzewo predyktorów skupień (CF). Wysoce zrównoważone drzewo przechowuje predyktory skupień dla grupowania hierarchicznego, w którym podobne rekordy stają się częścią węzłów tego samego drzewa.

- Liście drzewa CF są grupowane hierarchicznie w pamięci, aby wygenerować końcowy wynik grupowania. Najlepsza liczba skupień jest ustalana automatycznie. Jeśli określisz maksymalną liczbę skupień, wówczas zostanie ustalona najlepsza liczba skupień w podanym limicie.
- Wynik grupowania zostanie udoskonalony w drugim kroku, w którym względem danych stosowany jest algorytm podobny do algorytmu K-średnie.

Opcje zmiennej Dwustopniowa w Netezza

Ustawiając opcje zmiennej, można określić, że używane będą ustawienia roli zmiennej, które są zdefiniowane w węzłach poprzedzających. Przypisania do zmiennej można również wykonać ręcznie.

Wybierz element. Wybierz tę opcję, aby wykorzystać ustawienia roli z wcześniejszego węzła typu (lub z karty Typy wcześniejszego węzła źródłowego). Ustawienia roli to między innymi zmienne przewidywane i predyktory.

Użyj niestandardowych przypisań. Opcję tę należy wybrać w celu ręcznego przypisania zmiennych przewidywanych, predyktorów oraz innych ról.

Zmienne. Aby ręcznie przypisać pozycje z tej listy do zmiennych ról po prawej stronie, należy użyć klawiszy strzałek. Ikony wskazują prawidłowe poziomy pomiar dla każdego pola roli.

Identyfikator rekordu. Zmienna, która będzie używana jako unikatowy identyfikator rekordu.

Predyktory (dane wejściowe). Wybierz jedną lub więcej zmiennych jako dane wejściowe dla predykcji.

Opcje budowania zmiennej Dwustopniowa w Netezza

Ustawienie opcji budowania umożliwia dostosowanie modelu do konkretnego celu.

Jeśli zamierzasz zbudować model z opcjami domyślnymi, kliknij opcję **Wykonaj**.

Miara odległości. Ten parametr definiuje metodę pomiaru odległości między punktami danych. Większe odległości oznaczają większe niepodobieństwa. Opcje są następujące:

- Logarytm wiarygodności.** Miara wiarygodności stosuje do zmiennych rozkład prawdopodobieństwa. Zakłada się, że zmienne ilościowe mają rozkład normalny, natomiast kategoryjne rozkład wielomianowy. Zakłada się, że wszystkie zmienne są niezależne.
- Euklidesowa.** Odległość euklidesowa jest odległością „w linii prostej” pomiędzy dwoma punktami danych.
- Znormalizowana euklidesowa.** Miara Znormalizowana euklidesowa jest podobna do miary euklidesowej, ale jest znormalizowana przez kwadrat odchylenia standardowego. Miara Znormalizowana euklidesowa, w przeciwieństwie do miary euklidesowej, jest również skalo-niezmiennicza.

Numer skupienia. Ten parametr definiuje liczbę skupień do utworzenia. Opcje są następujące:

- Automatycznie wylicz liczbę grup.** Liczba skupień jest obliczana automatycznie. Maksymalną liczbę skupień można określić w polu **Maksimum**.
- Określ liczbę grup.** Określ liczbę skupień do utworzenia.

Statystyki. Ten parametr definiuje liczbę statystyk uwzględnionych w modelu. Opcje są następujące:

- Wszystko.** Uwzględnione są wszystkie statystyki dotyczące kolumny i wszystkie statystyki dotyczące wartości.

Uwaga: Ten parametr zawiera maksymalną liczbę statystyk i może wpłynąć na wydajność systemu. Jeśli wyświetlanie modelu w formacie graficznym nie jest wymagane, należy określić wartość **Brak**.

- Kolumny.** Uwzględnione są statystyki dotyczące kolumny.
- Brak.** Uwzględnione są tylko statystyki, które są wymagane do oceny modelu.

Replikacja wyników. Zaznacz to pole wyboru, jeśli zamierzasz ustawić wartość początkową generatora liczb losowych, aby replikować wyniki. Możesz określić wartość całkowitą lub utworzyć pseudolosową wartość całkowitą, klikając opcję **Utwórz**.

Netezza PCA

Analiza głównych składników (PCA, ang. Principal Component Analysis) to potężna technika redukcji danych przeznaczona do zmniejszania złożoności danych. PCA znajduje kombinacje liniowe zmiennych wejściowych, które umożliwiają określenie wariancji w całym zestawie zmiennych, pod warunkiem że składowe są zlokalizowane ortogonalnie (nie są skorelowane) do siebie. Celem jest znalezienie niewielkiej liczby zmiennych pochodnych (składników głównych) w efektywny sposób podsumowujących informacje w oryginalnym zestawie zmiennych wejściowych.

Opcje zmiennej PCA Netezza

Na karcie Zmienne można zdecydować, czy mają zostać użyte ustawienia roli zmiennej już zdefiniowane w węzłach poprzedzających, czy też przypisania zmiennych mają zostać dokonane ręcznie.

Użyj wstępnie zdefiniowanych ról. Ta opcja korzysta z ustawień roli (zmienne przewidywane, predyktory itd.) z poprzedzającego węzła Typ (lub z karty Typy poprzedzającego węzła źródłowego).

Użyj niestandardowych przypisań. Opcję tę należy wybrać w celu ręcznego przypisania zmiennych przewidywanych, predyktorów oraz innych ról na tym ekranie.

Zmienne. Aby ręcznie przypisać pozycje z tej listy do różnych zmiennych ról po prawej stronie ekranu, należy użyć klawiszy strzałek. Ikony wskazują prawidłowe poziomy pomiaru dla każdego pola roli.

Kliknij przycisk **Wszystkie**, aby wybrać wszystkie pola na liście, lub kliknij przycisk poziomy pojedynczego pomiaru, aby wybrać wszystkie pola, które posiadają ten poziom pomiaru.

Identyfikator rekordu. Zmienna, która będzie używana jako unikatowy identyfikator rekordu.

Predyktory (dane wejściowe). Wybierz jedną lub więcej zmiennych jako dane wejściowe dla predykcji.

Opcje budowania PCA Netezza

Na karcie Opcje budowania można ustawić wszystkie opcje budowania modelu. Można oczywiście po prostu kliknąć przycisk **Uruchom** w celu zbudowania modelu ze wszystkimi opcjami domyślnymi, lecz w normalnej sytuacji użytkownik potrzebuje zwykle dostosować proces budowy do swoich celów.

Wyśrodkuj dane przed obliczeniem PCA. Jeśli ta opcja jest zaznaczona (domyślnie jest), wówczas wykonuje wyśrodkowanie danych (znane również jako „odejmowanie średniej”) przed analizą. Wyśrodkowanie danych jest niezbędne do zapewnienia, że pierwszy składnik główny opisuje kierunek wariancji maksymalnej. W przeciwnym wypadku ten składnik mógłby bliżej odpowiadać średniej z danych. Normalnie zaznaczenie tej opcji można usunąć w celu poprawy wydajności, jeśli dane zostały już przygotowane w ten sposób.

Przeprowadź skalowanie danych przed obliczeniem PCA. Wybranie tej opcji powoduje skalowanie danych przed analizą. Taki sposób postępowania sprawia, że analiza jest mniej dowolna, gdy różne zmienne są mierzone w różnych jednostkach. W najprostszej postaci skalowanie danych można osiągnąć poprzez podział każdej zmiennej przez jej wariancję standardową.

W celu obliczenia PCA użyj mniej dokładnej, ale szybszej metody. Wybranie tej opcji powoduje, że algorytm stosuje mniej dokładną, ale szybszą metodę (forceEigensolve) znajdowania składników głównych.

Zarządzanie modelami IBM Netezza Analytics

Modele IBM Netezza Analytics są dodawane do obszaru roboczego i palety Modele tak samo, jak inne modele IBM SPSS Modeler, a ponadto mogą być używane w bardzo podobny sposób. Istnieje jednak kilka istotnych różnic, ponieważ każdy model IBM Netezza Analytics utworzony w produkcie IBM SPSS Modeler w rzeczywistości odwołuje się do modelu zapisanego na serwerze bazy danych. Z tego względu strumień może działać poprawnie, pod warunkiem że nawiąże połączenie z bazą danych, w której model został utworzony, przy czym tabela modelu nie może zostać zmieniona przez proces zewnętrzny.

Ocenianie modeli IBM Netezza Analytics

W obszarze roboczym modele reprezentuje ikona złotego modelu użytkowego. Głównym przeznaczeniem modelu użytkowego jest dokonywanie oceny danych w celu wygenerowania predykcji lub umożliwienie przeprowadzenia dalszej analizy właściwości modelu. Oceny są dodawane w postaci jednego lub większej liczby dodatkowych pól danych, które można uwidocznic poprzez dołączenie węzła tabeli do modelu użytkowego i uruchomienie tej gałęzi strumienia, co zostało opisane w dalszej części niniejszej sekcji. W oknach dialogowych niektórych modeli użytkowych, np. tych dla drzewa decyzyjnego lub drzewa regresji, dodatkowo istnieje karta Model, na której dostępna jest wizualna reprezentacja modelu.

Dodatkowe zmienne wyróżnia przedrostek $\$<id>$ dodany do nazwy zmiennej przewidywanej, przy czym $<id>$ jest zależne od modelu i identyfikuje typ dodawanych informacji. Różne identyfikatory zostały opisane w tematach poświęconych poszczególnym modelom użytkowym.

Aby wyświetlić oceny, wykonaj poniższe czynności:

1. Dołącz węzeł tabeli do modelu użytkowego.
2. Otwórz węzeł tabeli.
3. Kliknij przycisk **Uruchom**.
4. Przewiń w prawo w oknie wyników tabeli, aby wyświetlić dodatkowe zmienne i ich oceny.

Karta Serwer modelu użytkowego Netezza

Na karcie Serwer można ustawić opcje serwera na potrzeby oceniania modelu. Można również w dalszym ciągu używać połączenia z serwerem, które zostało określone wcześniej, lub przenieść dane do innej bazy danych, którą można określić w tym miejscu.

Szczegóły serwera DB Netezza. W tym miejscu należy określić szczegóły połączenia z bazą danych, która będzie używana na potrzeby modelu.

- **Użyj połączenia zdefiniowanego w strumieniu.** (ustawienie domyślne) Wybranie tej opcji powoduje użycie szczegółów połączenia określonych w węźle poprzedzającym, na przykład węźle źródła bazy danych. *Uwaga:* ta opcja działa tylko wówczas, gdy wszystkie węzły poprzedzające mogą wykorzystywać funkcję analizy wstępnej SQL. W takim przypadku nie ma potrzeby przesuwać danych poza bazę danych, ponieważ SQL w pełni implementuje wszystkie węzły poprzedzające.
- **Przenieś dane do bazy przez połączenie.** Przenosi dane do bazy danych określonej w tym miejscu. Dzięki temu modelowanie może działać, jeśli dane znajdują się w innej bazie danych IBM Netezza lub baza pochodzi od innego dostawcy, a także w przypadku, gdy dane znajdują się w pliku płaskim. Ponadto dane są przenoszone z powrotem do bazy danych określonej w tej opcji, jeśli dane zostały wyodrębnione, ponieważ węzeł nie zrealizował analizy wstępnej SQL. Kliknij przycisk **Edycja**, aby przeglądać w poszukiwaniu połączenia i wybrać połączenie. *Ostrzeżenie:* produkt IBM Netezza Analytics jest zwykle używany z bardzo dużymi zestawami danych. Przesyłanie dużych ilości danych między bazami danych lub z bazy i do bazy może być bardzo czasochłonne i w miarę możliwości należy tego unikać.

Nazwa modelu. Nazwa modelu. Nazwa jest przedstawiona wyłącznie w celach informacyjnych; w tym miejscu nie można jej zmienić.

Modele użytkowe drzewa decyzyjnego Netezza

Model użytkowy drzewa decyzyjnego przedstawia wynik operacji modelowania i umożliwia również ustawienie niektórych opcji oceniania modelu.

Gdy uruchamiasz strumień zawierający model użytkowy drzewa decyzyjnego, domyślnie węzeł dodaje jedną nową zmienną, której nazwa wywodzi się z nazwy zmiennej przewidywanej.

Tabela 16. Zmienna oceny modelu dla drzewa decyzyjnego.

Nazwa dodanej zmiennej	Znaczenie
\$I-nazwa_zmiennej	Przewidywana wartość dla bieżącego rekordu.

Jeśli wybierzesz opcję **Oblicz prawdopodobieństwa przypisanych klas dla rekordów oceniania** w węźle modelowania lub modelu użytkowym, a następnie uruchomisz strumień, zostanie dodana dalsza zmienna.

Tabela 17. Zmienna oceny modelu dla drzewa decyzyjnego — informacje dodatkowe.

Nazwa dodanej zmiennej	Znaczenie
\$IP-nazwa_zmiennej	Wartość ufności (od 0,0 do 1,0) dla predykcji.

Model użytkowy drzewa decyzyjnego Netezza — karta Model

Karta **Model** przedstawia ważność predyktorów modelu drzewa decyzyjnego w formacie graficznym. Długość słupka reprezentuje ważność predyktora.

Uwaga: Gdy pracujesz z produktem IBM Netezza Analytics Version 2.x lub z jego wcześniejszą wersją, zawartość drzewa decyzyjnego jest przedstawiana tylko w formacie tekstowym.

W przypadku tych wersji wyświetlane są następujące informacje:

- Każdy wiersz tekstu odpowiada węzłowi lub liściowi.
- Wcięcie odzwierciedla poziom drzewa.
- W przypadku węzła wyświetlany jest warunek podziału.
- W przypadku liścia widoczna jest przypisana etykieta klasy.

Model użytkowy drzewa decyzyjnego Netezza — karta Ustawienia

Karta Ustawienia umożliwia ustawienie niektórych opcji oceny modelu.

Uwzględnij zmienne wejściowe. Wybranie tej opcji powoduje przekazanie wszystkich oryginalnych zmiennych wejściowych do dalszych etapów i dołączenie dodatkowej zmiennej lub zmiennych modelowania do każdego wiersza danych. Jeśli zaznaczenie tego pola wyboru zostanie usunięte, wówczas zostanie przekazana tylko zmienna ID rekordu i dodatkowe zmienne modelowania, dzięki czemu strumień będzie wykonywany szybciej.

Oblicz prawdopodobieństwa przypisanych klas dla rekordów oceniania. (tylko Drzewo decyzyjne i Naive Bayes) Wybranie tej opcji oznacza, że dodatkowe zmienne modelowania zawierają zmienną ufności (tj. prawdopodobieństwo), a także zmienną predykcji. Jeśli zaznaczenie tego pola wyboru zostanie usunięte, wówczas wygenerowana zostanie tylko zmienna predykcji.

Użyj deterministycznych danych wejściowych. Jeśli ta opcja jest wybrana, zapewnia, że dowolny algorytm Netezza, który wykonuje wiele przebiegów względem tego samego widoku, wykorzysta ten sam zestaw danych dla każdego przebiegu. Jeśli zaznaczenie tego pola wyboru zostanie usunięte, aby przedstawić, że używane są dane niedeterministyczne, wówczas utworzona zostanie tabela tymczasowa, w której przechowywane będą dane wynikowe przetwarzania, np. wygenerowane przez węzeł podziału; ta tabela zostanie usunięta po utworzeniu modelu.

Model użytkowy drzewa decyzyjnego Netezza — karta Przeglądarka

Karta **Przeglądarka** przedstawia reprezentację drzewa modelu drzewa w taki sam sposób, w jaki produkt SPSS Modeler przedstawia model drzewa decyzyjnego.

Uwaga: Jeśli model został zbudowany w produkcie IBM Netezza Analytics Version 2.x lub z jego wcześniejszej wersji, wówczas karta **Przeglądarka** jest pusta.

Modele użytkowe K-średnie Netezza

Model użytkowy K-średnie zawiera wszystkie informacje przechwycone przez model skupień, a także informacje o danych uczących i procesie estymacji.

Po uruchomieniu strumienia zawierającego model użytkowy K-średnie węzeł dodaje dwie nowe zmienne zawierające przynależność do skupień i odległość od przypisanego środka skupienia dla tego rekordu. Nowa zmienna o nazwie \$KM-K-średnie dotyczy przynależności do skupienia, a nowa zmienna o nazwie \$KMD-K-średnie dotyczy odległości od środka skupienia.

Modele użytkowe K-średnie Netezza — karta Model

Karta **Model** zawiera różne widoki graficzne, które przedstawiają statystyki podsumowujące i rozkłady dla zmiennych skupień. Dane można wyeksportować z modelu lub można wyeksportować widok jako grafikę.

Gdy pracujesz z produktem IBM Netezza Analytics Version 2.x lub z jego wcześniejszą wersją, lub jeśli zbudujesz model z Mahalanobis jako miarą odległości, wówczas zawartość modeli K-średnie jest przedstawiana tylko w formacie tekstowym.

W przypadku tych wersji wyświetlane są następujące informacje:

- **Statystyki podsumowujące.** W przypadku skupienia najmniejszego i największego statystyki podsumowujące przedstawiają liczbę rekordów. Statystyki podsumowujące przedstawiają także procent zestawu danych, który jest przyjmowany przez te skupienia. Lista przedstawia także stosunek rozmiaru największego skupienia do najmniejszego.
- **Podsumowanie grupowania.** Podsumowanie grupowania przedstawia listę skupień, które są tworzone przez algorytm. W przypadku każdego skupienia tabela przedstawia liczbę rekordów w danym skupieniu, a także średnią odległość od środka skupienia dla tych rekordów.

Modele użytkowe K-średnie Netezza — karta Ustawienia

Karta Ustawienia umożliwia ustawienie niektórych opcji oceny modelu.

Uwzględnij zmienne wejściowe. Wybranie tej opcji powoduje przekazanie wszystkich oryginalnych zmiennych wejściowych do dalszych etapów i dołączenie dodatkowej zmiennej lub zmiennych modelowania do każdego wiersza danych. Jeśli zaznaczenie tego pola wyboru zostanie usunięte, wówczas zostanie przekazana tylko zmienna ID rekordu i dodatkowe zmienne modelowania, dzięki czemu strumień będzie wykonywany szybciej.

Miara odległości. Metoda używana w celu pomiaru odległości między punktami danych; większa odległość oznacza większe niepodobieństwo. Opcje są następujące:

- **Euklidesowa.** (opcja domyślna) Odległość między dwoma punktami wyliczona jako długość łączącej je linii prostej.
- **Manhattan.** Odległość między dwoma punktami jest sumą bezwzględnych różnic między ich współrzędnymi.
- **Canberra.** Podobna do odległości Manhattan, ale bardziej czuła w przypadku punktów danych położonych bliżej początku układu współrzędnych.
- **Maksimum.** Odległość między dwoma punktami jest obliczana jako największa z różnic wzdłuż dowolnej współrzędnej.

Modele użytkowe Sieci Bayesa Netezza

Model użytkowy Sieci Bayesa udostępnia sposób ustawiania opcji na potrzeby oceny modelu.

Gdy uruchamiasz strumień zawierający model użytkowy Sieci Bayesa, węzeł dodaje jedną nową zmienną, której nazwa wywodzi się z nazwy zmiennej przewidywanej.

Tabela 18. Zmienna oceny modelu dla Sieci Bayesa.

Nazwa dodanej zmiennej	Znaczenie
\$BN-nazwa_zmiennej	Przewidywana wartość dla bieżącego rekordu.

Dodatkową zmienną można zobaczyć, dołączając węzeł Tabela do modelu użytkowego i uruchamiając węzeł Tabela.

Model użytkowy Sieci Bayesa Netezza — karta Ustawienia

Na karcie Ustawienia można ustawić opcje na potrzeby oceniania modelu.

Zmienna przewidywana. Jeśli wymagana jest ocena zmiennej przewidywanej, która różni się od bieżącej zmiennej, wybierz nową zmienną w tym miejscu.

Identyfikator rekordu. Jeśli nie jest określona żadna zmienna ID rekordu, wybierz w tym miejscu zmienną do użycia.

Typ predykcji. Zmienność algorytmu predykcji wybranego do użycia:

- **Najlepszy (najbardziej skorelowany sąsiad).** (opcja domyślna) Używa najbardziej skorelowanego węzła sąsiedniego.
- **Sąsiedzi (ważona predykcja sąsiadów).** Stosuje predykcję ważoną względem wszystkich węzłów sąsiednich.
- **'Sąsiedzi NN (sąsiedzie inni niż null).** Podobnie jak poprzednia opcja, z wyjątkiem tego, że ignoruje węzły z wartościami null (czyli węzły odpowiadające atrybutom, które mają brakujące wartości dla instancji, względem której predykcja jest obliczana).

Uwzględnij zmienne wejściowe. Wybranie tej opcji powoduje przekazanie wszystkich oryginalnych zmiennych wejściowych do dalszych etapów i dołączenie dodatkowej zmiennej lub zmiennych modelowania do każdego wiersza danych. Jeśli zaznaczenie tego pola wyboru zostanie usunięte, wówczas zostanie przekazana tylko zmienna ID rekordu i dodatkowe zmienne modelowania, dzięki czemu strumień będzie wykonywany szybciej.

Modele użytkowe Netezza Naive Bayes

Model użytkowy Naive Bayes udostępnia sposób ustawiania opcji na potrzeby oceny modelu.

Gdy uruchamiasz strumień zawierający model użytkowy Naive Bayes, domyślnie węzeł dodaje jedną nową zmienną, której nazwa wywodzi się z nazwy zmiennej przewidywanej.

Tabela 19. Zmienna oceny modelu dla Naive Bayes — ustawienia domyślne.

Nazwa dodanej zmiennej	Znaczenie
\$I-nazwa_zmiennej	Przewidywana wartość dla bieżącego rekordu.

Jeśli wybierzesz opcję **Oblicz prawdopodobieństwa przypisanych klas dla rekordów oceniania** w węźle modelowania lub modelu użytkowym, a następnie uruchomisz strumień, zostaną dodane kolejne dwie zmienne.

Tabela 20. Zmienne oceny modelu dla Naive Bayes — dodatkowe.

Nazwa dodanej zmiennej	Znaczenie
\$IP-nazwa_zmiennej	Numerator Bayesowski klasy dla instancji (czyli produkt prawdopodobieństwa klasy poprzedniej i prawdopodobieństwa warunkowego wartości atrybutu instancji).
\$ILP-nazwa_zmiennej	Logarytm naturalny tej drugiej części.

Dodatkowe zmienne można zobaczyć, dołączając węzeł Tabela do modelu użytkowego i uruchamiając węzeł Tabela.

Modele użytkowe Netezza Naive Bayes — karta Ustawienia

Na karcie Ustawienia można ustawić opcje na potrzeby oceniania modelu.

Uwzględnij zmienne wejściowe. Wybranie tej opcji powoduje przekazanie wszystkich oryginalnych zmiennych wejściowych do dalszych etapów i dołączenie dodatkowej zmiennej lub zmiennych modelowania do każdego wiersza danych. Jeśli zaznaczenie tego pola wyboru zostanie usunięte, wówczas zostanie przekazana tylko zmienna ID rekordu i dodatkowe zmienne modelowania, dzięki czemu strumień będzie wykonywany szybciej.

Oblicz prawdopodobieństwa przypisanych klas dla rekordów oceniania. (tylko Drzewo decyzyjne i Naive Bayes) Wybranie tej opcji oznacza, że dodatkowe zmienne modelowania zawierają zmienną ufności (tj. prawdopodobieństwo), a także zmienną predykcji. Jeśli zaznaczenie tego pola wyboru zostanie usunięte, wówczas wygenerowana zostanie tylko zmienna predykcji.

Popraw dokładność prawdopodobieństwa dla małych lub niezrównoważonych zbiorów danych. Podczas obliczania prawdopodobieństw ta opcja wywołuje technikę estymacji m , która unika prawdopodobieństw zerowych. Ten rodzaj estymacji prawdopodobieństw może działać wolniej, ale może zwracać lepsze wyniki w przypadku małych lub znacznie niezrównoważonych zbiorów danych.

Modele użytkowe KNN Netezza

Model użytkowy KNN udostępni sposób ustawiania opcji na potrzeby oceny modelu.

Gdy uruchamiasz strumień zawierający model użytkowy KNN, węzeł dodaje jedną nową zmienną, której nazwa wywodzi się z nazwy zmiennej przewidywanej.

Tabela 21. Zmienna oceny modelu dla KNN.

Nazwa dodanej zmiennej	Znaczenie
\$KNN-nazwa_zmiennej	Przewidywana wartość dla bieżącego rekordu.

Dodatkową zmienną można zobaczyć, dołączając węzeł Tabela do modelu użytkowego i uruchamiając węzeł Tabela.

Model użytkowy KNN Netezza — karta Ustawienia

Na karcie Ustawienia można ustawić opcje na potrzeby oceniania modelu.

Miara odległości. Metoda używana w celu pomiaru odległości między punktami danych; większa odległość oznacza większe niepodobieństwo. Opcje są następujące:

- **Euklidesowa.** (opcja domyślna) Odległość między dwoma punktami wyliczona jako długość łączącej je linii prostej.
- **Manhattan.** Odległość między dwoma punktami jest sumą bezwzględnych różnic między ich współrzędnymi.
- **Canberra.** Podobna do odległości Manhattan, ale bardziej czuła w przypadku punktów danych położonych bliżej początku układu współrzędnych.
- **Maksimum.** Odległość między dwoma punktami jest obliczana jako największa z różnic wzdłuż dowolnej współrzędnej.

Liczba najbliższych sąsiadów (k). Liczba najbliższych sąsiadów dla konkretnej obserwacji. Należy pamiętać, że większa liczba obserwacji najbliższego sąsiedztwa nie zawsze oznacza dokładniejszy model.

Wybranie opcji k zapewnia równowagę między zapobieganiem przeuczeniu (to może być szczególnie istotne w przypadku zaszumionych danych) i rozdzielaniem (uzyskiwaniem różnych predykcji w przypadku podobnych instancji). Zwykle konieczne jest dostosowanie wartości k dla każdego zestawu danych, przy czym typowe wartości należą do zakresu od 1 do kilku tuzinów.

Uwzględnij zmienne wejściowe. Wybranie tej opcji powoduje przekazanie wszystkich oryginalnych zmiennych wejściowych do dalszych etapów i dołączenie dodatkowej zmiennej lub zmiennych modelowania do każdego wiersza

danych. Jeśli zaznaczenie tego pola wyboru zostanie usunięte, wówczas zostanie przekazana tylko zmienna ID rekordu i dodatkowe zmienne modelowania, dzięki czemu strumień będzie wykonywany szybciej.

Standaryzacja pomiarów przed obliczeniem. Wybranie tej opcji powoduje standaryzację pomiarów dla ciągłych danych wejściowych przed obliczeniem wartości odległości.

Użyj zestawów podstawowych w celu poprawy wydajności w przypadku dużych zestawów danych. Wybranie tej opcji powoduje używanie próbkowania zestawu podstawowego w celu przyspieszenia obliczeń, gdy używane są duże zestawy danych.

Modele użytkowe grupowania dzielącego Netezza

Model użytkowy Grupowanie dzielące udostępnia sposób ustawiania opcji na potrzeby oceny modelu.

Gdy uruchamiasz strumień zawierający model użytkowy Grupowanie dzielące, węzeł dodaje dwie nowe zmienne, której nazwy wywodzą się od nazwy zmiennej przewidywanej.

Tabela 22. Zmienna oceny modelu dla Grupowania dzielącego.

Nazwa dodanej zmiennej	Znaczenie
\$DC-nazwa_zmiennej	Identyfikator podgrupy, do której przypisany jest bieżący rekord.
\$DCD-nazwa_zmiennej	Odległość od środka podgrupy dla bieżącego rekordu.

Dodatkowe zmienne można zobaczyć, dołączając węzeł Tabela do modelu użytkowego i uruchamiając węzeł Tabela.

Model użytkowy grupowania dzielącego Netezza — karta Ustawienia

Na karcie Ustawienia można ustawić opcje na potrzeby oceniania modelu.

Uwzględnij zmienne wejściowe. Wybranie tej opcji powoduje przekazanie wszystkich oryginalnych zmiennych wejściowych do dalszych etapów i dołączenie dodatkowej zmiennej lub zmiennych modelowania do każdego wiersza danych. Jeśli zaznaczenie tego pola wyboru zostanie usunięte, wówczas zostanie przekazana tylko zmienna ID rekordu i dodatkowe zmienne modelowania, dzięki czemu strumień będzie wykonywany szybciej.

Miara odległości. Metoda używana w celu pomiaru odległości między punktami danych; większa odległość oznacza większe niepodobieństwo. Opcje są następujące:

- **Euklidesowa.** (opcja domyślna) Odległość między dwoma punktami wyliczona jako długość łączącej je linii prostej.
- **Manhattan.** Odległość między dwoma punktami jest sumą bezwzględnych różnic między ich współrzędnymi.
- **Canberra.** Podobna do odległości Manhattan, ale bardziej czuła w przypadku punktów danych położonych bliżej początku układu współrzędnych.
- **Maksimum.** Odległość między dwoma punktami jest obliczana jako największa z różnic wzdłuż dowolnej współrzędnej.

Poziom stosowanej hierarchii. Poziom hierarchii, jaki powinien być stosowany względem danych.

Modele użytkowe PCA Netezza

Model użytkowy PCA udostępnia sposób ustawiania opcji na potrzeby oceny modelu.

Gdy uruchamiasz strumień zawierający model użytkowy PCA, domyślnie węzeł dodaje jedną nową zmienną, której nazwa wywodzi się z nazwy zmiennej przewidywanej.

Tabela 23. Zmienna oceny modelu dla PCA.

Nazwa dodanej zmiennej	Znaczenie
\$F-nazwa_zmiennej	Przewidywana wartość dla bieżącego rekordu.

Jeśli określisz wartość większą niż 1 w polu **Liczba składników głównych...** w węźle modelowania lub modelu użytkowym, a następnie uruchomisz strumień, węzeł doda nową zmienną dla każdego składnika. W tym przypadku nazwy zmiennych będą poprzedzone przedrostkiem *n*, gdzie *n* jest numerem składnika. Na przykład, jeśli model ma nazwę *pca* i zawiera trzy składniki, wówczas nowe zmienne będą miały następujące nazwy: *\$F-pca-1*, *\$F-pca-2* i *\$F-pca-3*.

Dodatkowe zmienne można zobaczyć, dołączając węzeł Tabela do modelu użytkowego i uruchamiając węzeł Tabela.

Model użytkowy PCA Netezza — karta Ustawienia

Na karcie Ustawienia można ustawić opcje na potrzeby oceniania modelu.

Liczba podstawowych składników do użycia w odwzorowaniu. Liczba podstawowych składników, do których zamierzasz zredukować zestaw danych. Wartość nie może przekroczyć liczby atrybutów (zmiennych wejściowych).

Uwzględnij zmienne wejściowe. Wybranie tej opcji powoduje przekazanie wszystkich oryginalnych zmiennych wejściowych do dalszych etapów i dołączenie dodatkowej zmiennej lub zmiennych modelowania do każdego wiersza danych. Jeśli zaznaczenie tego pola wyboru zostanie usunięte, wówczas zostanie przekazana tylko zmienna ID rekordu i dodatkowe zmienne modelowania, dzięki czemu strumień będzie wykonywany szybciej.

Modele użytkowe drzewa regresji Netezza

Model użytkowy drzewa regresji udostępnia sposób ustawiania opcji na potrzeby oceny modelu.

Gdy uruchamiasz strumień zawierający model użytkowy drzewa regresji, domyślnie węzeł dodaje jedną nową zmienną, której nazwa wywodzi się z nazwy zmiennej przewidywanej.

Tabela 24. Zmienna oceny modelu dla Drzewa regresji.

Nazwa dodanej zmiennej	Znaczenie
<i>\$I-nazwa_zmiennej</i>	Przewidywana wartość dla bieżącego rekordu.

Jeśli wybierzesz opcję **Oblicz oszacowaną wariancję** w węźle modelowania lub modelu użytkowym, a następnie uruchomisz strumień, zostanie dodana dalsza zmienna.

Tabela 25. Zmienna oceny modelu dla Drzewa regresji — dodatkowe.

Nazwa dodanej zmiennej	Znaczenie
<i>\$IV-nazwa_zmiennej</i>	Oszacowane wariancje wartości przewidywanej.

Dodatkowe zmienne można zobaczyć, dołączając węzeł Tabela do modelu użytkowego i uruchamiając węzeł Tabela.

Model użytkowy drzewa regresji Netezza — karta Model

Karta **Model** przedstawia ważność predyktorów modelu drzewa regresji w formacie graficznym. Długość słupka reprezentuje ważność predyktora.

Uwaga: Gdy pracujesz z produktem IBM Netezza Analytics Version 2.x lub z jego wcześniejszą wersją, zawartość drzewa regresji jest przedstawiana tylko w formacie tekstowym.

W przypadku tych wersji wyświetlane są następujące informacje:

- Każdy wiersz tekstu odpowiada węzłowi lub liściowi.
- Wcięcie odzwierciedla poziom drzewa.
- W przypadku węzła wyświetlany jest warunek podziału.
- W przypadku liścia widoczna jest przypisana etykieta klasy.

Model użytkowy Drzewo regresji Netezza — karta Ustawienia

Na karcie Ustawienia można ustawić opcje na potrzeby oceniania modelu.

Uwzględnij zmienne wejściowe. Wybranie tej opcji powoduje przekazanie wszystkich oryginalnych zmiennych wejściowych do dalszych etapów i dołączenie dodatkowej zmiennej lub zmiennych modelowania do każdego wiersza danych. Jeśli zaznaczenie tego pola wyboru zostanie usunięte, wówczas zostanie przekazana tylko zmienna ID rekordu i dodatkowe zmienne modelowania, dzięki czemu strumień będzie wykonywany szybciej.

Oblicz oszacowaną wariancję. Wskazuje, czy wariancja przypisanych klas powinna być uwzględniona w wynikach.

Model użytkowy Drzewo regresji Netezza — karta Przeglądarka

Karta **Przeglądarka** przedstawia reprezentację drzewa modelu drzewa w taki sam sposób, w jaki produkt SPSS Modeler przedstawia model drzewa regresji.

Uwaga: Jeśli model został zbudowany w produkcie IBM Netezza Analytics Version 2.x lub z jego wcześniejszej wersji, wówczas karta **Przeglądarka** jest pusta.

Modele użytkowe regresji liniowej Netezza

Model użytkowy regresji liniowej udostępnia sposób ustawiania opcji na potrzeby oceny modelu.

Gdy uruchamiasz strumień zawierający model użytkowy regresji liniowej, węzeł dodaje jedną nową zmienną, której nazwa wywodzi się z nazwy zmiennej przewidywanej.

Tabela 26. Zmienna oceny modelu dla regresji liniowej.

Nazwa dodanej zmiennej	Znaczenie
\$LR-nazwa_zmiennej	Przewidywana wartość dla bieżącego rekordu.

Model użytkowy Regresja liniowa Netezza — karta Ustawienia

Na karcie Ustawienia można ustawić opcje na potrzeby oceniania modelu.

Uwzględnij zmienne wejściowe. Wybranie tej opcji powoduje przekazanie wszystkich oryginalnych zmiennych wejściowych do dalszych etapów i dołączenie dodatkowej zmiennej lub zmiennych modelowania do każdego wiersza danych. Jeśli zaznaczenie tego pola wyboru zostanie usunięte, wówczas zostanie przekazana tylko zmienna ID rekordu i dodatkowe zmienne modelowania, dzięki czemu strumień będzie wykonywany szybciej.

Model użytkowy szeregów czasowych Netezza

Model użytkowy udostępnia wyniki operacji modelowania szeregu czasowego. Dane wyjściowe obejmują następujące zmienne.

Tabela 27. Zmienne wyjściowe modelu szeregów czasowych

Zmienna	Opis
TSID	Identyfikator szeregu czasowego; zawartość zmiennej określonej dla identyfikatora szeregu czasowego na karcie Zmienne węzła modelowania. Więcej informacji można znaleźć w temacie “Opcje zmiennych szeregu czasowego Netezza” na stronie 96.
CZAS	Okres w bieżącym szeregu czasowym.
HISTORY	Wartości z danych historycznych (tych, które są używane do przeprowadzenia predykcji). Ta zmienna jest uwzględniana tylko wówczas, gdy opcja W wyniku uwzględnij wartości historyczne jest zaznaczona na karcie Ustawienia w modelu użytkowym.
\$STS-INTERPOLATED	Wartości interpolowane, jeśli są używane. Ta zmienna jest uwzględniana tylko wówczas, gdy opcja W wyniku uwzględnij wartości interpolowane jest zaznaczona na karcie Ustawienia w modelu użytkowym. Interpolacja to opcja na karcie Opcje budowania w węźle modelowania.

Tabela 27. Zmienne wyjściowe modelu szeregów czasowych (kontynuacja)

Zmienna	Opis
\$STS-FORECAST	Wartości prognozy dla szeregu czasowego.

W celu wyświetlenia wyniku modelu należy dołączyć węzeł Tabela (z karty wyniku palety węzła) do modelu użytkowego, a następnie uruchomić węzeł Tabela.

Model użytkowy Netezza Time Series — karta Ustawienia

Na karcie Ustawienia można określić opcje w celu dostosowania wyniku modelu.

Nazwa modelu. Nazwa modelu określona na karcie Opcje modelu w węźle modelowania.

Pozostałe opcje są takie same, jak te na karcie Opcje modelowania w węźle modelowania.

Użytkowy uogólniony model liniowy Netezza

Model użytkowy udostępnia wyniki operacji modelowania.

Gdy uruchamiasz strumień zawierający użytkowy uogólniony model liniowy, węzeł dodaje jedną nową zmienną, której nazwa wywodzi się z nazwy zmiennej przewidywanej.

Tabela 28. Zmienna oceny modelu dla Uogólnione liniowe.

Nazwa dodanej zmiennej	Znaczenie
\$GLM-nazwa_zmiennej	Przewidywana wartość dla bieżącego rekordu.

Na karcie Model wyświetlane są różne statystyki dotyczące modelu.

Dane wyjściowe obejmują następujące zmienne.

Tabela 29. Zmienne wyjściowe z uogólnionego modelu liniowego.

Zmienna wyjściowa	Opis
Parametr	Parametry (czyli zmienne predyktora) używane przez model. Są to kolumny liczbowe i nominalne, a także wyraz wolny (składnik stały w modelu regresji).
Beta	Współczynnik korelacji (tj. korelacji liniowej modelu).
Błąd standardowy	Odchylenie standardowe dla beta.
Test	Statystyki testu używane w celu oceny poprawności parametru.
Wartość p	Prawdopodobieństwo błędu przy założeniu, że parametr jest istotny.
Podsumowanie reszt	
Typ reszty	Typ reszty predykcji, dla której pokazane są wartości podsumowujące.
RSS	Wartość reszty.
df	Stopnie swobody dla reszty.
Wartość p	Prawdopodobieństwo błędu. Wyższa wartość oznacza model słabo dopasowany; niska wartość wskazuje dobre dopasowanie.

Użytkowy uogólniony model liniowy Netezza — karta Ustawienia

Na karcie Ustawienia można dostosować wyniki modelu.

Opcja jest taka sama, jak pokazana dla Oceniania w węźle modelowania. Więcej informacji można znaleźć w temacie “Opcje uogólnionego modelu liniowego Netezza — opcje oceniania” na stronie 86.

Model użytkowy Dwustopniowa Netezza

Po uruchomieniu strumienia zawierającego model użytkowy Dwustopniowa węzeł dodaje dwie nowe zmienne zawierające przynależność do skupień i odległość od przypisanego środka skupienia dla tego rekordu. Nowa zmienna o nazwie \$TS-Dwustopniowa dotyczy przynależności do skupienia, a nowa zmienna o nazwie \$TSP-Dwustopniowa dotyczy odległości od środka skupienia.

Model użytkowy Dwustopniowa Netezza — karta Model

Karta **Model** zawiera różne widoki graficzne, które przedstawiają statystyki podsumowujące i rozkłady dla zmiennych skupień. Dane można wyeksportować z modelu lub można wyeksportować widok jako grafikę.

Rozdział 7. Modelowanie w bazie danych za pomocą produktu IBM DB2 for z/OS

IBM SPSS Modeler and IBM DB2 for z/OS

Produkt SPSS Modeler obsługuje integrację z produktem DB2 for z/OS, co umożliwia uruchamianie zaawansowanych analiz na serwerach DB2 for z/OS. Dostęp do tych funkcji jest możliwy za pośrednictwem interfejsu graficznego SPSS Modeler oraz zorientowanego na przepływ pracy środowiska programistycznego. Pozwala to uruchamiać algorytmy eksploracji danych bezpośrednio w środowisku DB2 for z/OS, z wykorzystaniem produktu IBM DB2 Analytics Accelerator.

Produkt SPSS Modeler obsługuje integrację następujących algorytmów z produktu DB2 for z/OS.

- Drzewa decyzyjne
- K-średnie
- Naive Bayes
- Drzewo regresji
- Dwustopniowa

Wymagania dotyczące integracji z produktem IBM DB2 for z/OS

Poniższe warunki stanowią warunki wstępne do przeprowadzenia modelowania w bazie danych z użyciem produktów DB2 for z/OS i IBM DB2 Analytics Accelerator for z/OS. W celu upewnienia się, że te warunki są spełnione, może być konieczna konsultacja z administratorem bazy danych.

- Produkt IBM SPSS Modeler uruchamiany w trybie lokalnym lub względem instalacji SPSS Modeler Server w systemie Windows lub UNIX
- DB2 for z/OS w wersji 10 lub nowszej oraz DB2 Analytics Accelerator for z/OS w wersji 5
- IBM SPSS Data Access Pack 7.1
- Jeden z następujących systemów na serwerze, na którym uruchamiany jest produkt SPSS Modeler Server:
 - IBM DB2 Data Server Driver for ODBC and CLI
 - Dowolna wersja produktu DB2 for Linux, UNIX oraz system operacyjny Windows ze źródłem danych ODBC skonfigurowanym na potrzeby produktu DB2 for z/OS
- Licencja na produkt DB2 Connect for System z
- Funkcje generowania i optymalizacji kodu SQL włączone w produkcie SPSS Modeler
- IBM SPSS Modeler Scoring Adapter for zEnterprise 17.0

Aktywacja integracji z produktem IBM DB2 Analytics Accelerator for z/OS

Aktywacja integracji z produktem DB2 Analytics Accelerator for z/OS składa się z następujących kroków:

- Konfigurowanie programów DB2 for z/OS i DB2 Analytics Accelerator for z/OS
- Utworzenie źródła ODBC
- Aktywacja integracji produktu IBM DB2 for z/OS w produkcie IBM SPSS Modeler
- Włączenie opcji generowania i optymalizacji kodu SQL w programie SPSS Modeler
- Aktywacja produktu IBM SPSS Modeler Server Scoring Adapter na potrzeby produktu DB2 for z/OS

Konfigurowanie programu IBM DB2 for z/OS i IBM Analytics Accelerator for z/OS

Sposób konfigurowania programów DB2 for z/OS oraz Analytics Accelerator for z/OS opisano w następujących serwisach WWW:

DB2 Analytics Accelerator for z/OS.

Tworzenie źródła ODBC dla produktów IBM DB2 for z/OS oraz IBM DB2 Analytics Accelerator

Więcej informacji na temat sposobu aktywacji połączenia między produktami DB2 for z/OS a IBM DB2 Analytics Accelerator można znaleźć w następujących serwisach WWW:

- W przypadku wersji 4: DB2 Analytics Accelerator for z/OS 4.1.0
- W przypadku wersji 3: DB2 Analytics Accelerator for z/OS 3.1.0
- Włączanie akceleracji zapytań za pomocą produktu IBM DB2 Analytics Accelerator for ODBC oraz aplikacji JDBC bez modyfikowania aplikacji
- Błąd SQL sterownika ODBC podczas uruchamiania zapytania w programie DB2 Analytics Accelerator for z/OS

Aktywacja integracji produktu IBM DB2 for z/OS w produkcie IBM SPSS Modeler

Aby aktywować integrację programu DB2 for DB2 w programie SPSS Modeler, wykonaj następujące kroki:

1. Otwórz plik `user.prefs` w następującym katalogu w systemie Windows:
`c:\Users\twoj_id_uzytkownika\AppData\Roaming\IBM\SPSS\Modeler\18\Defaults\`, gdzie *twoj_id_uzytkownika* to Twój identyfikator użytkownika.
Jeśli plik nie istnieje, musisz go utworzyć.
2. Dołącz następujący wiersz:
`EnableIDAA=true`.
3. W katalogu `config` programu SPSS Modeler otwórz plik `odbc-db2-accelerator-names.cfg`.
Jeśli plik nie istnieje, musisz go utworzyć.
4. Dodaj nazwy wszystkich źródeł danych oraz nazwy wszystkich akceleratorów, na przykład:
`dsn1, acceleratorname1`
`dsn2, acceleratorname2`
5. W menu głównym SPSS Modeler kliknij pozycje **Narzędzia > Opcje > Aplikacje pomocnicze**.
6. Kliknij kartę **IBM DB2 for z/OS**.
7. Wybierz opcję **Włącz integrację z IBM DB2 for z/OS Data Mining**, a następnie kliknij przycisk **OK**.

Informacje pokrewne:

“Tworzenie źródła ODBC dla produktów IBM DB2 for z/OS oraz IBM DB2 Analytics Accelerator”

Włączenie opcji generowania i optymalizacji kodu SQL

Istnieje możliwość pracy z bardzo dużymi zbiorami danych, dlatego ze względu na wydajność należy włączyć opcje generowania i optymalizacji kodu SQL w produkcie IBM SPSS Modeler.

Aby skonfigurować program SPSS Modeler, wykonaj poniższe czynności:

1. Z menu IBM SPSS Modeler wybierz pozycje **Narzędzia > Właściwości strumienia > Opcje**.
2. Kliknij opcję **Optymalizacja** w panelu nawigacji.
3. Upewnij się, że włączona jest opcja **Generuj kod SQL kierowany do bazy**. To ustawienie jest niezbędne, ponieważ zapewnia poprawne działanie modelowania w bazie danych.
4. Wybierz opcje **Optymalizuj operacje generujące kod SQL i Optymalizuj inne wykonywane operacje** (nie jest to ściśle wymagane, ale zdecydowanie zalecane w celu poprawy wydajności).

Budowanie modeli za pomocą produktu IBM DB2 for z/OS

Dla każdego obsługiwane algorytmu istnieje odpowiadający mu węzeł modelowania. Dostęp do węzłów modelowania DB2 for z/OS można uzyskiwać z karty Modelowanie w bazie na palecie węzłów.

Zagadnienia dotyczące danych

Pola w źródle danych zawierają zmienne różnych typów danych, co jest uzależnione od węzła modelowania. W produkcie SPSS Modeler typy danych są określane mianem **poziomów pomiaru**. W karcie Zmienne dla węzła modelowania używane są ikony oznaczające dozwolone typy poziomów pomiaru dla zmiennych wejściowych i przewidywanych.

Zmienna przewidywana. Zmienna przewidywana to zmienna, której wartość próbujemy przewidzieć. Gdy możliwe jest określenie zmiennej przewidywanej, wówczas tylko jedno pole danych źródłowych można wybrać jako zmienną przewidywaną.

Zmienna ID rekordów. Określa zmienną używaną w celu unikalnej identyfikacji każdej obserwacji. Na przykład może to być zmienna identyfikacyjna, taka jak *CustomerID*. Jeśli dane źródłowe nie zawierają zmiennej identyfikacyjnej, można utworzyć tę zmienną za pomocą węzła wyliczeń, co przedstawia poniższa procedura.

1. Wybierz węzeł źródłowy.
2. Na karcie Zmienne w palecie węzłów kliknij dwukrotnie węzeł wyliczeń.
3. Otwórz węzeł wyliczeń, klikając dwukrotnie jego ikonę na obszarze roboczym.
4. W polu **Zmienna wyliczana** wpisz (na przykład) ID.
5. W polu **Formuła** wpisz @INDEX i kliknij przycisk **OK**.
6. Połącz węzeł wyliczeń z pozostałą częścią strumienia.

Postępowanie z wartościami null

Jeśli wartości wejściowe zawierają wartości null, stosowanie niektórych węzłów DB2 for z/OS może spowodować komunikaty o błędach lub długo działające strumienie, dlatego zalecamy usunięcie rekordów zawierających wartości null. Należy użyć poniższej metody.

1. Dołącz węzeł selekcji do węzła źródłowego.
2. Ustaw opcję **Tryb** węzła selekcji na wartość **Odrzuć**.
3. Wprowadź poniższy łańcuch do pola **Warunek**:
`@NULL(field1) [or @NULL(field2)[... or @NULL(fieldN)]`
Upewnij się, że uwzględniona została każda zmienna wejściowa.
4. Połącz węzeł selekcji z pozostałą częścią strumienia.

Wyniki modelu

Strumień zawierający węzeł modelowania DB2 for z/OS może zwracać trochę inne wyniki za każdym razem, gdy jest uruchamiany. Dzieje się tak, ponieważ kolejność, w jakiej węzeł odczytuje źródło danych, nie jest zawsze taka sama, gdyż przed budowaniem modelu dane są wczytywane do tabel tymczasowych. Jednak różnice wywołane przez ten efekt są pomijalne.

Komentarze ogólne

- W produkcie SPSS Collaboration and Deployment Services nie jest możliwe tworzenie konfiguracji oceniania przy użyciu strumieni zawierających węzły modelowania DB2 for z/OS.
- W przypadku modelu utworzonych przez węzły DB2 for z/OS nie jest możliwy eksport ani import kodu PMML.

Modele IBM DB2 for z/OS — Opcje zmiennych

Na karcie Zmienne można zdecydować, czy mają zostać użyte ustawienia roli zmiennej już zdefiniowane w węzłach poprzedzających, czy też przypisania zmiennych mają zostać dokonane ręcznie.

Użyj wstępnie zdefiniowanych ról. Ta opcja korzysta z ustawień roli (zmienne przewidywane, predyktory itd.) z poprzedzającego węzła Typ (lub z karty Typy poprzedzającego węzła źródłowego).

Użyj niestandardowych przypisań. Opcję tę należy wybrać w celu ręcznego przypisania zmiennych przewidywanych, predyktorów oraz innych ról na tym ekranie.

Zmienne. Aby ręcznie przypisać pozycje z tej listy do różnych zmiennych ról po prawej stronie ekranu, należy użyć klawiszy strzałek. Ikony wskazują prawidłowe poziomy pomiaru dla każdego pola roli.

Kliknij przycisk **Wszystkie**, aby wybrać wszystkie pola na liście, lub kliknij przycisk poziomu pojedynczego pomiaru, aby wybrać wszystkie pola, które posiadają ten poziom pomiaru.

Zmienna przewidywana. Wybierz jedną zmienną jako zmienną przewidywaną dla predykcji. W przypadku algorytmu Uogólnione modele liniowe wybierz także zmienną **Próby** na tym ekranie.

Identyfikator rekordu. Zmienna, która będzie używana jako unikatowy identyfikator rekordu.

Predyktory (dane wejściowe). Wybierz jedną lub więcej zmiennych jako dane wejściowe dla predykcji.

Modele IBM DB2 for z/OS — Opcje serwera

Na karcie Serwer określ system DB2 for z/OS, w którym zostanie zbudowany model.

- **Użyj połączenia zdefiniowanego w strumieniu.** (ustawienie domyślne) Wybranie tej opcji powoduje użycie szczegółów połączenia określonych w węzle poprzedzającym, na przykład węzle źródła bazy danych. *Uwaga:* ta opcja działa tylko wówczas, gdy wszystkie węzły poprzedzające mogą wykorzystywać funkcję analizy wstępnej SQL. W takim przypadku nie ma potrzeby przesuwać danych poza bazę danych, ponieważ SQL w pełni implementuje wszystkie węzły poprzedzające.
- **Przenieś dane do bazy przez połączenie.** Przenosi dane do bazy danych określonej w tym miejscu. Dzięki temu modelowanie może działać, jeśli dane znajdują się w innej bazie danych IBM, lub baza pochodzi od innego dostawcy, a także w przypadku, gdy dane znajdują się w pliku płaskim. Ponadto dane są przenoszone z powrotem do bazy danych określonej w tej opcji, jeśli dane zostały wyodrębnione, ponieważ węzeł nie zrealizował analizy wstępnej SQL. Kliknij przycisk **Edycja**, aby przeglądać w poszukiwaniu połączenia i wybrać połączenie.

Komentarz:

Nazwa źródła danych ODBC jest osadzona w każdym strumieniu SPSS Modeler. Jeśli strumień utworzony w jednym hoście zostanie wykonany na innym hoście, wówczas nazwa źródła danych musi być taka sama na każdym hoście. Alternatywnie inne źródło danych można wybrać na karcie Serwer w każdym węzle źródłowym lub węzle modelowania.

Modele IBM DB2 for z/OS — Opcje modelu

Na karcie Opcje modelu można zdecydować o wyborze nazwy dla modelu lub o jej wygenerowaniu automatycznie.

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu, w przypadkach gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Zastąp istniejący, jeśli nazwa została użyta. Jeśli to pole wyboru zostanie zaznaczone, wówczas wszelkie istniejące modele o tej samej nazwie zostaną zastąpione.

Modele IBM DB2 for z/OS – K-średnie

Węzeł K-średnie implementuje algorytm *k*-średnie, który zapewnia metodę analizy skupień. Za pomocą tego węzła można grupować zestaw danych do odrębnych grup.

Ten algorytm jest algorytmem grupowania zależnym od odległości, którego działanie opiera się na metryce (funkcji) odległości w celu pomiaru podobieństwa między punktami danych. Punkty danych są przypisane do najbliższego skupienia zgodnie z używaną metryką odległości.

Algorytm działa, wykonując kilka iteracji tego samego podstawowego procesu, w którym każda instancja ucząca jest przypisywana do najbliższego skupienia (względem określonej funkcji odległości, stosowanej do instancji i centrum skupienia). Wszystkie centra skupień są następnie obliczane ponownie jako wektory wartości atrybutu średniej instancji przypisanych do poszczególnych skupień.

Modele IBM DB2 for z/OS — Opcje zmiennej K-średnie

Na karcie Zmienne można zdecydować, czy mają zostać użyte ustawienia roli zmiennej już zdefiniowane w węzłach poprzedzających, czy też przypisania zmiennych mają zostać dokonane ręcznie.

Użyj wstępnie zdefiniowanych ról. Ta opcja korzysta z ustawień roli (zmienne przewidywane, predyktory itd.) z poprzedzającego węzła Typ (lub z karty Typy poprzedzającego węzła źródłowego).

Użyj niestandardowych przypisań. Opcję tę należy wybrać w celu ręcznego przypisania zmiennych przewidywanych, predyktorów oraz innych ról na tym ekranie.

Zmienne. Aby ręcznie przypisać pozycje z tej listy do różnych zmiennych ról po prawej stronie ekranu, należy użyć klawiszy strzałek. Ikony wskazują prawidłowe poziomy pomiaru dla każdego pola roli.

Kliknij przycisk **Wszystkie**, aby wybrać wszystkie pola na liście, lub kliknij przycisk poziomu pojedynczego pomiaru, aby wybrać wszystkie pola, które posiadają ten poziom pomiaru.

Identyfikator rekordu. Zmienna, która będzie używana jako unikatowy identyfikator rekordu.

Predyktory (dane wejściowe). Wybierz jedną lub więcej zmiennych jako dane wejściowe dla predykcji.

Modele IBM DB2 for z/OS — Opcje budowania zmiennej K-średnie

Ustawienie opcji budowania umożliwi dostosowanie modelu do konkretnego celu.

Jeśli zamierzasz zbudować model z opcjami domyślnymi, kliknij opcję **Wykonaj**.

Miara odległości. Ten parametr definiuje metodę pomiaru odległości między punktami danych. Większe odległości oznaczają większe niepodobieństwa. Wybierz jedną z poniższych opcji:

- **Euklidesowa.** Odległość euklidesowa jest odległością „w linii prostej” pomiędzy dwoma punktami danych.
- **Znormalizowana euklidesowa.** Miara Znormalizowana euklidesowa jest podobna do miary euklidesowej, ale jest znormalizowana przez kwadrat odchylenia standardowego. Miara Znormalizowana euklidesowa, w przeciwieństwie do miary euklidesowej, jest również skalo-niezmiennicza.

Liczba grup. Ten parametr definiuje liczbę skupień do utworzenia.

Maksymalna liczba iteracji. Algorytm wykonuje kilka iteracji tego samego procesu. Ten parametr definiuje liczbę iteracji, po których szkolenie modelu jest zatrzymywane.

Statystyki. Ten parametr definiuje liczbę statystyk uwzględnionych w modelu. Wybierz jedną z poniższych opcji:

- **Wszystko.** Uwzględnione są wszystkie statystyki dotyczące kolumny i wszystkie statystyki dotyczące wartości.

Uwaga: Ten parametr zawiera maksymalną liczbę statystyk i może wpłynąć na wydajność systemu. Jeśli wyświetlanie modelu w formacie graficznym nie jest wymagane, należy określić wartość **Brak**.

- **Kolumny.** Uwzględnione są statystyki dotyczące kolumny.
- **Brak.** Uwzględnione są tylko statystyki, które są wymagane do oceny modelu.

Replikacja wyników. Zaznacz to pole wyboru, jeśli zamierzasz ustawić wartość początkową generatora liczb losowych, aby replikować wyniki. Możesz określić wartość całkowitą lub utworzyć pseudolosową wartość całkowitą, klikając opcję **Utwórz**.

Modele IBM DB2 for z/OS – Naive Bayes

Naive Bayes to dobrze znany algorytm stosowany w przypadku problemów z klasyfikacją. Model jest określany jako naiwny, ponieważ traktuje wszystkie proponowane zmienne predykcji jako niezależne od siebie. Naive Bayes to szybki, skalowalny algorytm, który oblicza prawdopodobieństwa warunkowe kombinacji atrybutów i atrybutu przewidywanego. Na podstawie danych uczących wyznaczane jest niezależne prawdopodobieństwo. To prawdopodobieństwo określa wiarygodność każdej klasy przewidywanej z uwzględnieniem wystąpienia każdej kategorii wartości z poszczególnych zmiennych wejściowych.

Modele IBM DB2 for z/OS – Drzewa decyzyjne

Drzewo decyzyjne jest strukturą hierarchiczną, która reprezentuje model klasyfikacji. Za pomocą modelu drzewa decyzyjnego można opracować system klasyfikacji, aby przewidywać lub klasyfikować przyszłe obserwacje z zestawu danych uczących. Klasyfikacja przyjmuje postawę struktury drzewa, w której gałęzie reprezentują punkty podziału w klasyfikacji. Podziały rekurencyjnie rozdzielają dane na podgrupy, aż do osiągnięcia punktu zatrzymania. Węzły drzewa w punktach zatrzymania są znane pod nazwą *liście*. Każdy liść przypisuje etykietę znaną jako *etykieta klasy* do członków jego podgrupy lub klasy.

Modele IBM DB2 for z/OS — Opcje zmiennych drzewa decyzyjnego

Na karcie Zmienne można zdecydować, czy mają zostać użyte ustawienia roli zmiennej już zdefiniowane w węzłach poprzedzających, czy też przypisania zmiennych mają zostać dokonane ręcznie.

Użyj wstępnie zdefiniowanych ról. Ta opcja korzysta z ustawień roli (zmienne przewidywane, predyktory itd.) z poprzedzającego węzła Typ (lub z karty Typy poprzedzającego węzła źródłowego).

Użyj niestandardowych przypisań. Opcję tę należy wybrać w celu ręcznego przypisania zmiennych przewidywanych, predyktorów oraz innych ról na tym ekranie.

Zmienne. Aby ręcznie przypisać pozycje z tej listy do różnych zmiennych ról po prawej stronie ekranu, należy użyć klawiszy strzałek. Ikony wskazują prawidłowe poziomy pomiaru dla każdego pola roli.

Kliknij przycisk **Wszystkie**, aby wybrać wszystkie pola na liście, lub kliknij przycisk poziomego pojedynczego pomiaru, aby wybrać wszystkie pola, które posiadają ten poziom pomiaru.

Zmienna przewidywana. Wybierz jedną zmienną jako zmienną przewidywaną dla predykcji.

Identyfikator rekordu. Zmienna, która będzie używana jako unikatowy identyfikator rekordu. Wartości tej zmiennej muszą być unikalne dla każdego rekordu, np. muszą to być identyfikatory klientów.

Waga instancji. Określenie zmiennej w tym miejscu umożliwia użycie wag instancji (jedna waga na jeden wiersz danych wejściowych) zamiast — lub dodatkowo — domyślnych wag klas (jedna waga na kategorię dla zmiennej przewidywanej). Zmienna, którą określa użytkownik w tym miejscu, musi zawierać wagę liczbową dla każdego wiersza danych wejściowych.

Predyktory (dane wejściowe). Umożliwia wybór zmiennej wejściowej lub kilku zmiennych. Działanie jest podobne jak w przypadku ustawienia roli zmiennej na *Zmienna wejściowa* w węźle Typ.

Modele IBM DB2 for z/OS — Opcje budowania drzewa decyzyjnego

Dla wzrostu drzewa dostępne są następujące opcje budowania:

Miara wzrostu. Te opcje kontrolują sposób wzrostu drzewa.

- **Miary zanieczyszczenia.** Ta miara ocenia najlepsze miejsce podziału drzewa. Jest to pomiar zmienności w podgrupie lub segmencie danych. Niski pomiar zanieczyszczenia wskazuje grupę, w której większość elementów ma podobne wartości dla kryterium lub zmiennej przewidywanej.

Obsługiwane pomiary są następujące: **Entropia i Gini**. Te miary bazują na prawdopodobieństwach członkostwa w kategorii dla gałęzi.

- **Maksymalna głębokość drzewa.** Maksymalna liczba liści, do której może rosnąć drzewo poniżej węzła głównego, czyli jest to liczba rekurencyjnych podziałów próby. Wartością domyślną tej właściwości jest 10, a wartość maksymalna, jaką można ustawić dla tej właściwości, wynosi 62.

Uwaga: Jeśli przeglądarka w modelu użytkowym przedstawia tekstową reprezentację modelu, wówczas wyświetlanych jest maksymalnie 12 poziomów drzewa.

Kryteria podziału. Te opcje kontrolują miejsce, w którym następuje zatrzymanie podziału drzewa.

- **Minimalne ulepszenie dla podziałów.** Minimalna wartość, o jaką musi zostać zmniejszone zanieczyszczenie, zanim w drzewie zostanie utworzony nowy podział. Celem budowy drzewa jest utworzenie podgrup o podobnych wartościach wyjściowych — czyli minimalizowanie zanieczyszczeń w ramach każdego węzła. Jeśli najlepszy podział dla gałęzi zmniejsza zanieczyszczenia o mniej, niż określono w kryteriach podziału, wówczas gałąź nie jest dzielona.
- **Minimalna liczba instancji dla podziału.** Minimalna liczba rekordów, jaka może zostać podzielona. Jeśli liczba pozostałych niepodzielonych rekordów jest mniejsza niż ta liczba, wówczas dalsze podziały nie są wykonywane. Tego pola można użyć, aby zapobiec tworzeniu małych podgrup w drzewie.

Statystyki. Ten parametr definiuje liczbę statystyk uwzględnionych w modelu. Wybierz jedną z poniższych opcji:

- **Wszystko.** Uwzględnione są wszystkie statystyki dotyczące kolumny i wszystkie statystyki dotyczące wartości.

Uwaga: Ten parametr zawiera maksymalną liczbę statystyk i może wpłynąć na wydajność systemu. Jeśli wyświetlanie modelu w formacie graficznym nie jest wymagane, należy określić wartość **Brak**.

- **Kolumny.** Uwzględnione są statystyki dotyczące kolumny.
- **Brak.** Uwzględnione są tylko statystyki, które są wymagane do oceny modelu.

Modele IBM DB2 for z/OS — Węzeł Drzewo decyzyjne – Wagi klas

W tym miejscu można przypisywać wagi do poszczególnych klas. Domyślnie wartość 1 jest przypisywana do wszystkich klas, co powoduje, że uzyskują równe wagi. Określenie różnych wag liczbowych dla różnych etykiet klas oznacza, że algorytm otrzymuje instrukcję odpowiedniego obciążenia zestawów uczących konkretnych klas.

W celu zmiany wagi kliknij ją dwukrotnie w kolumnie **Waga** i wprowadź żądane zmiany.

Wartość. Zestaw etykiet klas uzyskany na podstawie możliwych wartości zmiennej przewidywanej.

Waga. Waga, jaka zostanie przypisana do konkretnej klasy. Przypisanie wyższej wagi do klasy powoduje, że model staje się bardziej czuły dla tej klasy w porównaniu do innych klas.

Wagi klas można używać w połączeniu z wagami instancji.

Modele IBM DB2 for z/OS — Węzeł Drzewo decyzyjne – Przcycinanie drzewa

Za pomocą opcji przycinania można określić kryteria przycinania dla drzewa decyzyjnego. Celem przycinania jest zmniejszenie ryzyka przeuczenia poprzez usunięcie przerośniętych podgrup, które nie poprawiają oczekiwanej dokładności w przypadku nowych danych.

Miara przycinania. Domyślna miara przycinania — **Dokładność** — zapewnia, że oszacowana dokładność modelu mieści się w dozwolonych limitach po usunięciu liścia z drzewa. Jeśli podczas przycinania chcesz uwzględnić wagi klas, użyj alternatywnej opcji **Ważona dokładność**.

Dane dla przycinania. W celu oszacowania oczekiwanej dokładności nowych danych można użyć części lub wszystkich danych uczących. Alternatywnie do tego celu można użyć osobnego zestawu danych do przycinania z podanej tabeli.

- **Użyj wszystkich danych uczących.** Ta opcja (domyślna) powoduje, że wykorzystywane są wszystkie dane uczące w celu oszacowania dokładności modelu.
- **Użyj % danych uczących do przycinania.** Ta opcja umożliwia podział danych na dwa zestawy — jeden jest używany do nauki, a drugi do przycinania, przy czym określony w opcji procent dotyczy danych do przycinania.
- Jeśli chcesz określić wartość początkową generatora liczb losowych, aby upewnić się, że dane zostaną podzielone tak samo za każdym razem, gdy uruchomisz strumień, wybierz opcję **Replikuj wyniki**. Możesz określić wartość całkowitą w polu **Wartość startowa generatora użyta do przycinania** lub kliknąć opcję **Utwórz**, co spowoduje utworzenie pseudolosowej wartości całkowitej.
- **Użyj danych z istniejącej tabeli.** Określ nazwę tabeli osobnego zestawu danych do przycinania w celu oszacowania dokładności modelu. Takie postępowanie jest traktowane jako bardziej niezawodne niż korzystanie z danych uczących.

Modele IBM DB2 for z/OS — Drzewo regresji

Drzewo regresji to algorytm bazujący na drzewie, który wielokrotnie dzieli próbę obserwacji, aby uzyskać podzbiór tego samego rodzaju odpowiednio do wartości liczbowej zmiennej przewidywanej. Drzewa regresji, podobnie jak drzewa decyzyjne, dekomponują dane na podzbiory, w których liście drzewa odpowiadają wystarczająco małym lub wystarczająco jednostajnym podzbiорom. Podziały są wybierane w celu zmniejszenia rozproszenia przewidywanych wartości atrybutu, dzięki czemu te wartości mogą być odpowiednio dobrze przewidziane na podstawie wartości średnich na liściach.

Modele IBM DB2 for z/OS — Opcje budowania drzewa regresji – Wzrost drzewa

Dla wzrostu drzewa i przycinania drzewa można ustawić opcje budowania.

Dla wzrostu drzewa dostępne są następujące opcje budowania:

Maksymalna głębokość drzewa. Maksymalna liczba liści, do której może rosnać drzewo poniżej węzła głównego, czyli jest to liczba rekurencyjnych podziałów próby. Wartością domyślną jest 62, co jest maksymalną głębokością drzewa na potrzeby modelowania.

Uwaga: Jeśli przeglądarka w modelu użytkowym przedstawia tekstową reprezentację modelu, wówczas wyświetlanych jest maksymalnie 12 poziomów drzewa.

Kryteria podziału. Te opcje kontrolują miejsce, w którym następuje zatrzymanie podziału drzewa.

- **Ewaluacyjna miara podziału.** Miara ewaluacyjna klasy ocenia najlepsze miejsce podziału drzewa.

Uwaga: Aktualnie jedyną możliwą opcją jest wariancja.

- **Minimalne ulepszenie dla podziałów.** Minimalna wartość, o jaką musi zostać zmniejszone zanieczyszczenie, zanim w drzewie zostanie utworzony nowy podział. Celem budowy drzewa jest utworzenie podgrup o podobnych wartościach wyjściowych — czyli minimalizowanie zanieczyszczeń w ramach każdego węzła. Jeśli najlepszy podział dla gałęzi zmniejsza zanieczyszczenia o mniej, niż określono w kryteriach podziału, wówczas gałąź nie jest dzielona.
- **Minimalna liczba instancji dla podziału.** Minimalna liczba rekordów, jaka może zostać podzielona. Jeśli liczba pozostałych niepodzielonych rekordów jest mniejsza niż ta liczba, wówczas dalsze podziały nie są wykonywane. Tego pola można użyć, aby zapobiec tworzeniu małych podgrup w drzewie.

Statystyki. Ten parametr definiuje liczbę statystyk uwzględnionych w modelu. Wybierz jedną z poniższych opcji:

- **Wszystko.** Uwzględnione są wszystkie statystyki dotyczące kolumny i wszystkie statystyki dotyczące wartości.

Uwaga: Ten parametr zawiera maksymalną liczbę statystyk i może wpłynąć na wydajność systemu. Jeśli wyświetlanie modelu w formacie graficznym nie jest wymagane, należy określić wartość **Brak**.

- **Kolumny.** Uwzględnione są statystyki dotyczące kolumny.
- **Brak.** Uwzględnione są tylko statystyki, które są wymagane do oceny modelu.

Modele IBM DB2 for z/OS — Opcje budowania drzewa regresji — Przycinanie drzewa

Za pomocą opcji przycinania można określić kryteria przycinania dla drzewa regresji. Celem przycinania jest zmniejszenie ryzyka przeuczenia poprzez usunięcie przerośniętych podgrup, które nie poprawiają oczekiwanej dokładności w przypadku nowych danych.

Miara przycinania. Miara przycinania zapewnia, że oszacowana dokładność modelu mieści się w dozwolonych limitach po usunięciu liścia z drzewa. Wybierz jedną z poniższych miar.

- **mse.** Średni błąd kwadratowy — (domyślnie) mierzy bliskość dopasowanej linii do punktów danych.
- **r2.** R-kwadrat — mierzy proporcję zmienności w zmiennej zależnej wyjaśnionej przez model regresji.
- **Pearsona.** Współczynnik korelacji Pearsona — mierzy siłę zależności między zmiennymi zależnymi liniowo, które mają rozkład normalny.
- **Spearman.** Współczynnik korelacji Spearmana — wykrywa zależności nieliniowe, które wydają się słabe w przypadku uwzględnienia korelacji Pearsona, ale w rzeczywistości mogą być silne.

Dane dla przycinania. W celu oszacowania oczekiwanej dokładności nowych danych można użyć części lub wszystkich danych uczących. Alternatywnie do tego celu można użyć osobnego zestawu danych do przycinania z podanej tabeli.

- **Użyj wszystkich danych uczących.** Ta opcja (domyślna) powoduje, że wykorzystywane są wszystkie dane uczące w celu oszacowania dokładności modelu.
- **Użyj % danych uczących do przycinania.** Ta opcja umożliwia podział danych na dwa zestawy — jeden jest używany do nauki, a drugi do przycinania, przy czym określony w opcji procent dotyczy danych do przycinania. Jeśli chcesz określić wartość początkową generatora liczb losowych, aby upewnić się, że dane zostaną podzielone tak samo za każdym razem, gdy uruchomisz strumień, wybierz opcję **Replikuj wyniki**. Możesz określić wartość całkowitą w polu **Wartość startowa generatora użyta do przycinania** lub kliknąć opcję **Utwórz**, co spowoduje utworzenie pseudolosowej wartości całkowitej.
- **Użyj danych z istniejącej tabeli.** Określ nazwę tabeli osobnego zestawu danych do przycinania w celu oszacowania dokładności modelu. Takie postępowanie jest traktowane jako bardziej niezawodne niż korzystanie z danych uczących.

Modele IBM DB2 for z/OS – Dwustopniowa

Węzeł Dwustopniowa implementuje algorytm dwustopniowy, który udostępnia metodę grupowania danych z dużych zestawów danych.

Za pomocą tego węzła można grupować dane, gdy uwzględniane są dostępne zasoby, np. ograniczona ilość pamięci i czasu.

Algorytm Dwustopniowa jest algorytmem eksploracji w bazie danych, który grupuje dane w następujący sposób:

1. Tworzone jest drzewo predyktorów skupień (CF). Wysoce zrównoważone drzewo przechowuje predyktory skupień dla grupowania hierarchicznego, w którym podobne rekordy stają się częścią węzłów tego samego drzewa.
2. Liście drzewa CF są grupowane hierarchicznie w pamięci, aby wygenerować końcowy wynik grupowania. Najlepsza liczba skupień jest ustalana automatycznie. Jeśli określisz maksymalną liczbę skupień, wówczas zostanie ustalona najlepsza liczba skupień w podanym limicie.
3. Wynik grupowania zostanie udoskonalony w drugim kroku, w którym względem danych stosowany jest algorytm podobny do algorytmu K-średnie.

Modele IBM DB2 for z/OS — Opcje zmiennej Dwustopniowa

Ustawiając opcje zmiennej, można określić, że używane będą ustawienia roli zmiennej, które są zdefiniowane w węzłach poprzedzających. Przypisania do zmiennej można również wykonać ręcznie.

Wybierz element. Wybierz tę opcję, aby wykorzystać ustawienia roli z wcześniejszego węzła typu (lub z karty Typy wcześniejszego węzła źródłowego). Ustawienia roli to między innymi zmienne przewidywane i predyktory.

Użyj niestandardowych przypisań. Opcję tę należy wybrać w celu ręcznego przypisania zmiennych przewidywanych, predyktorów oraz innych ról.

Zmienne. Aby ręcznie przypisać pozycje z tej listy do zmiennych ról po prawej stronie, należy użyć klawiszy strzałek. Ikony wskazują prawidłowe poziomy pomiar dla każdego pola roli.

Identyfikator rekordu. Zmienna, która będzie używana jako unikatowy identyfikator rekordu.

Predyktory (dane wejściowe). Wybierz jedną lub więcej zmiennych jako dane wejściowe dla predykcji.

Modele IBM DB2 for z/OS — Opcje budowania zmiennej Dwustopniowa

Ustawienie opcji budowania umożliwi dostosowanie modelu do konkretnego celu.

Jeśli zamierzasz zbudować model z opcjami domyślnymi, kliknij opcję **Wykonaj**.

Miara odległości. Ten parametr definiuje metodę pomiaru odległości między punktami danych. Większe odległości oznaczają większe niepodobieństwa. Opcja jest następująca:

- **Logarytm wiarygodności.** Miara wiarygodności stosuje do zmiennych rozkład prawdopodobieństwa. Zakłada się, że zmienne ilościowe mają rozkład normalny, natomiast kategoryjne rozkład wielomianowy. Zakłada się, że wszystkie zmienne są niezależne.

Numer skupienia. Ten parametr definiuje liczbę skupień do utworzenia. Opcje są następujące:

- **Automatycznie wylicz liczbę grup.** Liczba skupień jest obliczana automatycznie. Maksymalną liczbę skupień można określić w polu **Maksimum**.
- **Określ liczbę grup.** Określ liczbę skupień do utworzenia.

Statystyki. Ten parametr definiuje liczbę statystyk uwzględnionych w modelu. Opcje są następujące:

- **Wszystko.** Uwzględnione są wszystkie statystyki dotyczące kolumny i wszystkie statystyki dotyczące wartości.

Uwaga: Ten parametr zawiera maksymalną liczbę statystyk i może wpłynąć na wydajność systemu. Jeśli wyświetlanie modelu w formacie graficznym nie jest wymagane, należy określić wartość **Brak**.

- **Kolumny.** Uwzględnione są statystyki dotyczące kolumny.
- **Brak.** Uwzględnione są tylko statystyki, które są wymagane do oceny modelu.

Replikacja wyników. Zaznacz to pole wyboru, jeśli zamierzasz ustawić wartość początkową generatora liczb losowych, aby replikować wyniki. Możesz określić wartość całkowitą lub utworzyć pseudolosową wartość całkowitą, klikając opcję **Utwórz**.

Modele IBM DB2 for z/OS – Model użytkowy Dwustopniowa – Karta Model

Karta **Model** zawiera różne widoki graficzne, które przedstawiają statystyki podsumowujące i rozkłady dla zmiennych skupień. Dane można wyeksportować z modelu lub można wyeksportować widok jako grafikę.

Zarządzanie modelami IBM DB2 for z/OS

Modele DB2 for z/OS są dodawane do obszaru roboczego i palety Modele tak samo, jak inne modele IBM SPSS Modeler, a ponadto mogą być używane w bardzo podobny sposób.

Aby ocenić dane bezpośrednio w programie DB2 for z/OS, wykonaj następujące kroki:

1. Zainstaluj produkt SPSS Scoring Adapter w bazie danych DB2 for z/OS, w której znajdują się dane.
2. Upewnij się, że strumień ma połączenie z bazą danych DB2 for z/OS, w której znajdują się dane.

Ocenianie modeli IBM DB2 for z/OS

W obszarze roboczym modele reprezentuje ikona złotego modelu użytkowego. Głównym przeznaczeniem modelu użytkowego jest dokonywanie oceny danych w celu wygenerowania predykcji lub umożliwienie przeprowadzenia dalszej analizy właściwości modelu. Oceny są dodawane w postaci jednego lub większej liczby dodatkowych pól danych, które można uwidocznic poprzez dołączenie węzła tabeli do modelu użytkowego i uruchomienie tej gałęzi strumienia, co zostało opisane w dalszej części niniejszej sekcji. W oknach dialogowych niektórych modeli użytkowych, np. tych dla drzewa decyzyjnego lub drzewa regresji, dodatkowo istnieje karta Model, na której dostępna jest wizualna reprezentacja modelu.

Dodatkowe zmienne wyróżnia przedrostek $\$$ <id>- dodany do nazwy zmiennej przewidywanej, przy czym <id> jest zależne od modelu i identyfikuje typ dodawanych informacji. Różne identyfikatory zostały opisane w tematach poświęconych poszczególnym modelom użytkowym.

Aby wyświetlić oceny, wykonaj poniższe czynności:

1. Dołącz węzeł tabeli do modelu użytkowego.
2. Otwórz węzeł tabeli.
3. Kliknij przycisk **Uruchom**.
4. Przewiń w prawo w oknie wyników tabeli, aby wyświetlić dodatkowe zmienne i ich oceny.

Uwaga: Proces oceniania nie jest realizowany w akceleratorze, lecz w programie DB2 i w efekcie wymaga fizycznej lokalizacji tabeli wejściowej w DB2. Stąd jako dane wejściowe oceniania może być użyta tylko tabela oparta na bazie danych DB2 lub tabela akcelerowana. Jeśli strumień używa tabeli tylko akceleratora, występuje następujący błąd: "THE STATEMENT CANNOT BE EXECUTED BY DB2 OR IN THE ACCELERATOR".

Modele użytkowe drzewa decyzyjnego IBM DB2 for z/OS

Model użytkowy drzewa decyzyjnego przedstawia wynik operacji modelowania i umożliwia również ustawienie niektórych opcji oceniania modelu.

Gdy uruchamiasz strumień zawierający model użytkowy drzewa decyzyjnego, węzeł dodaje dwie nowe zmienne, której nazwy wywodzą się od nazwy zmiennej przewidywanej.

Tabela 30. Zmienna oceny modelu dla drzewa decyzyjnego.

Nazwa dodanej zmiennej	Znaczenie
$\$I$ -nazwa_zmiennej	Przewidywana wartość dla bieżącego rekordu.

Tabela 30. Zmienna oceny modelu dla drzewa decyzyjnego (kontynuacja).

Nazwa dodanej zmiennej	Znaczenie
\$IP-nazwa_zmiennej	Wartość ufności (od 0,0 do 1,0) dla predykcji.

Uwaga: Ze względu na obowiązujące w programie DB2 for z/OS ograniczenia nazwy kolumn mogą zostać obcięte.

Model użytkowy drzewa decyzyjnego IBM DB2 for z/OS – Karta Model

Karta **Model** przedstawia ważność predyktorów modelu drzewa decyzyjnego w formacie graficznym. Długość słupka reprezentuje ważność predyktora.

Model użytkowy drzewa decyzyjnego IBM DB2 for z/OS – Karta Przeglądarka

Karta **Przeglądarka** przedstawia reprezentację drzewa modelu drzewa w taki sam sposób, w jaki produkt SPSS Modeler przedstawia model drzewa decyzyjnego.

Model użytkowy K-średnie IBM DB2 for z/OS

Model użytkowy K-średnie zawiera wszystkie informacje przechwycone przez model skupień, a także informacje o danych uczących i procesie estymacji.

Po uruchomieniu strumienia zawierającego model użytkowy K-średnie węzeł dodaje dwie nowe zmienne zawierające przynależność do skupień i odległość od przypisanego środka skupienia dla tego rekordu. Nazwy nowych zmiennych pochodzą od nazwy modelu i mają prefiksy: \$KM- dotyczący przynależności do skupienia, oraz \$KMD- dotyczący odległości od środka skupienia. Na przykład, jeśli model ma nazwę Kmeans, nowe zmienne będą miały nazwy: \$KM-Kmeans i \$KMD-Kmeans.

Uwaga: Ze względu na obowiązujące w programie DB2 for z/OS ograniczenia nazwy kolumn mogą zostać obcięte.

Model użytkowy K-średnie IBM DB2 for z/OS – Karta Model

Karta **Model** zawiera różne widoki graficzne, które przedstawiają statystyki podsumowujące i rozkłady dla zmiennych skupień. Dane można wyeksportować z modelu lub można wyeksportować widok jako grafikę.

Modele użytkowe Naive Bayes IBM DB2 for z/OS

Gdy uruchamiasz strumień zawierający model użytkowy Naive Bayes, węzeł dodaje dwie nowe zmienne, której nazwy wywodzą się od nazwy zmiennej przewidywanej.

Tabela 31. Zmienna oceny modelu dla Naive Bayes.

Nazwa dodanej zmiennej	Znaczenie
\$I-nazwa_zmiennej	Przewidywana wartość dla bieżącego rekordu.
\$IP-nazwa_zmiennej	Wartość ufności (od 0,0 do 1,0) dla predykcji.

Uwaga: Ze względu na obowiązujące w programie DB2 for z/OS ograniczenia nazwy kolumn mogą zostać obcięte.

Dodatkowe zmienne można zobaczyć, dołączając węzeł Tabela do modelu użytkowego i uruchamiając węzeł Tabela.

Model użytkowy drzewa regresji IBM DB2 for z/OS

Gdy uruchamiasz strumień zawierający model użytkowy Drzewo regresji, węzeł dodaje dwie nowe zmienne, której nazwy wywodzą się od nazwy zmiennej przewidywanej.

Tabela 32. Zmienna oceny modelu dla Drzewa regresji.

Nazwa dodanej zmiennej	Znaczenie
\$I-nazwa_zmiennej	Przewidywana wartość dla bieżącego rekordu.

Tabela 32. Zmienna oceny modelu dla Drzewa regresji (kontynuacja).

Nazwa dodanej zmiennej	Znaczenie
\$IS-nazwa_zmiennej	Oszacowane odchylenia standardowe od wartości przewidywanej.

Uwaga: Ze względu na obowiązujące w programie DB2 for z/OS ograniczenia nazwy kolumn mogą zostać obcięte.

Dodatkowe zmienne można zobaczyć, dołączając węzeł Tabela do modelu użytkowego i uruchamiając węzeł Tabela.

Model użytkowy drzewa regresji IBM DB2 for z/OS — Karta Model

Karta **Model** przedstawia ważność predyktorów modelu drzewa regresji w formacie graficznym. Długość słupka reprezentuje ważność predyktora.

Model użytkowy drzewa regresji IBM DB2 for z/OS – Karta Przeglądarka

Karta **Przeglądarka** przedstawia reprezentację drzewa modelu drzewa w taki sam sposób, w jaki produkt SPSS Modeler przedstawia model drzewa regresji.

Model użytkowy Dwustopniowa IBM DB2 for z/OS

Po uruchomieniu strumienia zawierającego model użytkowy Dwustopniowa węzeł dodaje dwie nowe zmienne zawierające przynależność do skupień i odległość od przypisanego środka skupienia dla tego rekordu. Nazwy nowych zmiennych pochodzą od nazwy modelu i mają prefiksy: \$TS- dotyczący przynależności do skupienia, oraz \$TSD- dotyczący odległości od środka skupienia. Na przykład, jeśli model ma nazwę MDL, nowe zmienne będą miały nazwy: \$TS-MDL i \$TSD-MDL.

Uwagi

Niniejsza publikacja została przygotowana z myślą o produktach i usługach oferowanych w Stanach Zjednoczonych. Materiał ten jest również dostępny w IBM w innych językach. Jednakże w celu uzyskania dostępu do takiego materiału istnieje konieczność posiadania egzemplarza produktu w takim języku.

Produktów, usług lub opcji opisywanych w tym dokumencie IBM nie musi oferować we wszystkich krajach. Informacje o produktach i usługach dostępnych w danym kraju można uzyskać od lokalnego przedstawiciela IBM. Odwołanie do produktu, programu lub usługi IBM nie oznacza, że można użyć wyłącznie tego produktu, programu lub usługi IBM. Zamiast nich można zastosować ich odpowiednik funkcjonalny pod warunkiem, że nie narusza to praw własności intelektualnej IBM. Jednakże cała odpowiedzialność za ocenę przydatności i sprawdzenie działania produktu, programu lub usługi pochodzących od producenta innego niż IBM spoczywa na użytkowniku.

IBM może posiadać patenty lub złożone wnioski patentowe na towary i usługi, o których mowa w niniejszej publikacji. Przedstawienie niniejszej publikacji nie daje żadnych uprawnień licencyjnych do tychże patentów. Pisemne zapytania w sprawie licencji można przysyłać na adres:

*IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
U.S.A.*

Zapytania dotyczące zestawów znaków dwubajtowych (DBCS) należy kierować do lokalnych działów własności intelektualnej IBM (IBM Intellectual Property Department) lub wysłać je na piśmie na adres:

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

INTERNATIONAL BUSINESS MACHINES CORPORATION DOSTARCZA TĘ PUBLIKACJĘ W STANIE, W JAKIM SIĘ ZNAJDUJE ("AS IS") BEZ UDZIELANIA JAKICHKOLWIEK GWARANCJI (RĘKOJMIĘ RÓWNIEŻ WYŁĄCZA SIĘ), WYRAŹNYCH LUB DOMNIEMANYCH, A W SZCZEGÓLNOŚCI DOMNIEMANYCH GWARANCJI PRZYDATNOŚCI HANDLOWEJ, PRZYDATNOŚCI DO OKREŚLONEGO CELU ORAZ GWARANCJI, ŻE PUBLIKACJA TA NIE NARUSZA PRAW OSÓB TRZECICH. Ustawodawstwa niektórych krajów nie dopuszczają zastrzeżeń dotyczących gwarancji wyraźnych lub domniemanych w odniesieniu do pewnych transakcji; w takiej sytuacji powyższe zdanie nie ma zastosowania.

Informacje zawarte w niniejszej publikacji mogą zawierać nieścisłości techniczne lub błędy drukarskie. Informacje te są okresowo aktualizowane, a zmiany te zostaną uwzględnione w kolejnych wydaniach tej publikacji. IBM zastrzega sobie prawo do wprowadzania ulepszeń i/lub zmian w produktach i/lub programach opisanych w tej publikacji w dowolnym czasie, bez wcześniejszego powiadomienia.

Wszelkie wzmianki w tej publikacji na temat stron internetowych innych podmiotów zostały wprowadzone wyłącznie dla wygody użytkownika i w żadnym wypadku nie stanowią zachęty do ich odwiedzania. Materiały dostępne na tych stronach nie są częścią materiałów opracowanych dla tego produktu IBM, a użytkownik korzysta z nich na własną odpowiedzialność.

IBM ma prawo do używania i rozpowszechniania informacji przysłanych przez użytkownika w dowolny sposób, jaki uzna za właściwy, bez żadnych zobowiązań wobec ich autora.

Licencjodawcy tego programu, którzy chcieliby uzyskać informacje na temat programu w celu: (i) wdrożenia wymiany informacji między niezależnie utworzonymi programami i innymi programami (łącznie z tym opisywanym) oraz (ii) wspólnego wykorzystywania wymienianych informacji, powinni skontaktować się z:

*IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
U.S.A.*

Informacje takie mogą być udostępnione, o ile spełnione zostaną odpowiednie warunki, w tym, w niektórych przypadkach, zostanie uiszczona stosowna opłata.

Licencjonowany program opisany w niniejszej publikacji oraz wszystkie inne licencjonowane materiały dostępne dla tego programu są dostarczane przez IBM na warunkach określonych w Umowie IBM z Klientem, Międzynarodowej Umowie Licencyjnej IBM na Program lub w innych podobnych umowach zawartych między IBM i użytkownikami.

Dane dotyczące wydajności i cytowane przykłady zostały przedstawione jedynie w celu zobrazowania sytuacji. Faktyczne wyniki dotyczące wydajności mogą się różnić w zależności do konkretnych warunków konfiguracyjnych i operacyjnych.

Informacje dotyczące produktów innych podmiotów niż IBM zostały uzyskane od dostawców tych produktów, z ich publicznych ogłoszeń lub innych dostępnych publicznie źródeł. IBM nie testował tych produktów i nie może potwierdzić dokładności pomiarów wydajności, kompatybilności ani żadnych innych danych związanych z tymi produktami. Pytania dotyczące możliwości produktów innych podmiotów należy kierować do dostawców tych produktów.

Wszelkie stwierdzenia dotyczące przyszłych kierunków rozwoju i zamierzeń IBM mogą zostać zmienione lub wycofane bez powiadomienia.

Publikacja ta zawiera przykładowe dane i raporty używane w codziennej pracy. W celu kompleksowego ich zilustrowania, podane przykłady zawierają nazwiska osób prywatnych, nazwy przedsiębiorstw oraz nazwy produktów. Wszystkie te nazwy/nazwiska są fikcyjne i jakiegokolwiek podobieństwo do istniejących nazw/nazwisk jest całkowicie przypadkowe.

Znaki towarowe

IBM, logo IBM i ibm.com są znakami towarowymi lub zastrzeżonymi znakami towarowymi International Business Machines Corp. zarejestrowanymi w wielu systemach prawnych na całym świecie. Pozostałe nazwy produktów i usług mogą być znakami towarowymi IBM lub innych przedsiębiorstw. Aktualna lista znaków towarowych IBM dostępna jest w serwisie WWW IBM, w sekcji "Copyright and trademark information" (Informacje o prawach autorskich i znakach towarowych), pod adresem www.ibm.com/legal/copytrade.shtml.

Adobe, logo Adobe, PostScript oraz logo PostScript są znakami towarowymi lub zastrzeżonymi znakami towarowymi Adobe Systems Incorporated w Stanach Zjednoczonych i/lub w innych krajach.

Intel, logo Intel, Intel Inside, logo Intel Inside, Intel Centrino, logo Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium i Pentium są znakami towarowymi lub zastrzeżonymi znakami towarowymi Intel Corporation lub przedsiębiorstw podporządkowanych Intel Corporation w Stanach Zjednoczonych i w innych krajach.

Linux jest zastrzeżonym znakiem towarowym Linusa Torvaldsa w Stanach Zjednoczonych i/lub w innych krajach.

Microsoft, Windows, Windows NT oraz logo Windows są znakami towarowymi Microsoft Corporation w Stanach Zjednoczonych i/lub w innych krajach.

UNIX jest zastrzeżonym znakiem towarowym The Open Group w Stanach Zjednoczonych i/lub w innych krajach.

Java oraz wszystkie znaki towarowe i logo dotyczące języka Java są znakami towarowymi lub zastrzeżonymi znakami towarowymi Oracle i/lub przedsiębiorstw afiliowanych.

Warunki dotyczące dokumentacji produktu

Zezwolenie na korzystanie z tych publikacji jest przyznawane na poniższych warunkach.

Zakres stosowania

Niniejsze warunki stanowią uzupełnienie warunków używania serwisu WWW IBM.

Użytek osobisty

Użytkownik ma prawo kopiować te publikacje do własnego, niekomercyjnego użytku pod warunkiem zachowania wszelkich uwag dotyczących praw własności. Użytkownik nie ma prawa dystrybuować ani wyświetlać tych publikacji czy ich części, ani też wykonywać na ich podstawie prac pochodnych bez wyraźnej zgody IBM.

Użytek służbowy

Użytkownik ma prawo kopiować te publikacje, dystrybuować je i wyświetlać wyłącznie w ramach przedsiębiorstwa Użytkownika pod warunkiem zachowania wszelkich uwag dotyczących praw własności. Użytkownik nie ma prawa wykonywać na podstawie tych publikacji ani ich fragmentów prac pochodnych, kopiować ich, dystrybuować ani wyświetlać poza przedsiębiorstwem Użytkownika bez wyraźnej zgody IBM.

Prawa

Z wyjątkiem zezwoleń wyraźnie udzielonych w niniejszym dokumencie, nie udziela się jakichkolwiek innych zezwoleń, licencji ani praw, wyraźnych czy domniemanych, odnoszących się do tych publikacji czy jakichkolwiek informacji, danych, oprogramowania lub innej własności intelektualnej, o których mowa w niniejszym dokumencie.

IBM zastrzega sobie prawo do anulowania zezwolenia przyznanego w niniejszym dokumencie w każdej sytuacji, gdy, według uznania IBM, korzystanie z tych publikacji jest szkodliwe dla IBM lub jeśli IBM uzna, że warunki niniejszego dokumentu nie są przestrzegane.

Użytkownik ma prawo pobierać, eksportować lub reeksportować niniejsze informacje pod warunkiem zachowania bezwzględnej i pełnej zgodności z obowiązującym prawem i przepisami, w tym ze wszelkimi prawami i przepisami eksportowymi Stanów Zjednoczonych.

IBM NIE UDZIELA JAKICHKOLWIEK GWARANCJI, W TYM TAKŻE RĘKOJMI, DOTYCZĄCYCH TREŚCI TYCH PUBLIKACJI. PUBLIKACJE TE SĄ DOSTARCZANE W STANIE, W JAKIM SIĘ ZNAJDUJĄ ("AS-IS") BEZ UDZIELANIA JAKICHKOLWIEK GWARANCJI (RĘKOJMIĘ RÓWNIEŻ WYŁĄCZA SIĘ), WYRAŹNYCH CZY DOMNIEMANYCH, A W SZCZEGÓLNOŚCI DOMNIEMANYCH GWARANCJI PRZYDATNOŚCI HANDLOWEJ, PRZYDATNOŚCI DO OKREŚLONEGO CELU CZY NIENARUSZANIA PRAW OSÓB TRZECICH.

Indeks

A

- AI (ang. Attribute Importance)
 - Oracle Data Mining 43, 44
- Algorytm gaussowski
 - Oracle Support Vector Machine 33
- analiza spektralna, IBM Netezza Analytics 94
- Apriori
 - Microsoft 18
 - Oracle Data Mining 41, 42

B

- baza danych
 - modelowanie w bazie danych 8, 11, 13, 15, 21
 - modelowanie w bazie danych da ISW 50

C

- czynnik złożoności
 - Oracle Support Vector Machine 34

D

- dane tabelaryczne
 - węzeł ISW Association 58
- dane transakcyjne
 - węzeł ISW Association 58
- DB2
 - zarządzanie modelami 54
- dekompozycja trendu sezonowego, IBM Netezza Analytics 94
- dokumentacja 3
- drzewa decyzyjne
 - Microsoft Analysis Services 11, 13, 21
 - ocenie — opcje podsumowania 22
 - ocenie — opcje serwera 21
 - opcje modelu 17
 - opcje serwera 17
 - opcje zaawansowane 17
- drzewa regresji
 - IBM DB2 for z/OS 120, 121, 124, 125
 - IBM Netezza Analytics 81, 82, 108, 109
- drzewo decyzyjne
 - IBM DB2 for z/OS 118, 119, 120, 123, 124, 125
 - IBM Netezza Analytics 87, 88, 89, 103, 104, 109
 - Oracle Data Mining 37, 38
- DSN
 - konfigurowanie 13
- Dwustopniowa
 - IBM DB2 for z/OS 121, 122, 123, 125
 - IBM Netezza Analytics 99, 100, 111
- dzielenie danych 41

E

- edytor kategorii
 - węzeł ISW Association 61
- eksploracja 25, 48, 72
- eksploracja bazy danych
 - budowanie modeli 8
 - konfiguracja 13
 - opcje optymalizacji 8
 - przy użyciu produktu IBM SPSS Modeler 7
 - przygotowanie danych 8
 - przykład 24, 71
- eksport
 - Modele DB2 55
- Modeler usług Analysis Services 24
- epsilon
 - Oracle Support Vector Machine 34
- etykieta klasy, w modelach drzewa Netezza 87, 118

F

- funkcja distance
 - Oracle K-średnie 39

G

- generowanie kodu SQL 8
- generowanie węzłów 24
- grupowanie
 - IBM Netezza Analytics 107
 - InfoSphere Warehouse Data Mining 65
 - ocenie — opcje podsumowania 22
 - ocenie — opcje serwera 21
 - opcje modelu 17
 - opcje serwera 17
 - opcje zaawansowane 17
- grupowanie dzielące
 - IBM Netezza Analytics 82, 83
- Grupowanie dzielące
 - IBM Netezza Analytics 107
- grupowanie sekwencyjne
 - opcje modelu 17
- grupowanie sekwencyjne (Microsoft) 20
 - opcje zaawansowane 21
 - opcje zmiennych 20

I

- IBM
 - modelowanie asocjacyjne 49
 - modelowanie drzewa decyzyjnego 49
 - modelowanie grupowania demograficznego 49
 - modelowanie grupowania Kohonena 49
 - modelowanie Naive Bayes 49
 - modelowanie regresji 49
 - modelowanie regresji liniowej 49
 - modelowanie regresji logistycznej 49

IBM (kontynuacja)

- modelowanie regresji wielomianowej 49
- modelowanie sekwencji 49
- modelowanie szeregów czasowych 49
- zarządzanie modelami 54, 80
- IBM DB2 for z/OS 113
 - Drzewa decyzyjne 118
 - Drzewo regresji 120
 - Dwustopniowa 121
 - integracja z produktem IBM DB2 Analytics Accelerator for z/OS 113
 - K-średnie 117
 - konfigurowanie programu IBM DB2 for z/OS i IBM Analytics Accelerator for z/OS 114
 - konfigurowanie za pomocą produktu IBM SPSS Modeler 115, 116
 - model użytkowy drzewa regresji 124, 125
 - model użytkowy Dwustopniowa 123, 125
 - model użytkowy Naive Bayes 124
 - modele użytkowe drzewa decyzyjnego 123, 124, 125
 - Naive Bayes 118
 - opcje budowania drzewa decyzyjnego 119, 120
 - opcje budowania drzewa regresji Netezza 120, 121
 - opcje budowania K-średnie 117
 - opcje budowania zmiennej Dwustopniowa 122
 - opcje modelu 116
 - Opcje zmiennej Dwustopniowa 122
 - Opcje zmiennej K-średnie 117
 - opcje zmiennych 116
 - Opcje zmiennych drzewa decyzyjnego 118
 - Wartościowa informacja z modelu K-średnie 124
 - wymagania dotyczące integracji z produktem IBM DB2 for z/OS 113
 - zarządzanie modelami DB2 for z/OS 123
- IBM Netezza Analytics 75
 - Drzewa decyzyjne 87
 - Drzewo regresji 81
 - Dwustopniowa 99
 - Grupowanie dzielące 82
 - K-średnie 92
 - konfigurowanie za pomocą produktu IBM SPSS Modeler 75, 76, 78, 79
 - model użytkowy drzewa regresji 108, 109
 - model użytkowy Dwustopniowa 111
 - model użytkowy grupowania dzielącego 107
 - model użytkowy KNN 106
 - model użytkowy Naive Bayes 105, 106
 - model użytkowy PCA 107, 108
 - model użytkowy regresji liniowej 109
 - Model użytkowy regresji liniowej 109
 - model użytkowy Sieci Bayesa 104, 105

IBM Netezza Analytics (*kontynuacja*)
 model użytkowy szeregów
 czasowych 109, 110
 modele użytkowe drzewa
 decyzyjnego 103, 104, 109
 Naive Bayes 93
 Najbliższe sąsiedztwo (KNN) 90
 opcje budowania drzewa decyzyjnego 88,
 89
 opcje budowania drzewa regresji
 Netezza 81, 82
 opcje budowania grupowania
 dzielącego 83
 opcje budowania K-średnie 92
 opcje budowania PCA 101
 opcje budowania regresji liniowej 90
 opcje budowania Sieci Bayesa 94
 opcje budowania szeregu czasowego 97,
 98
 opcje budowania zmiennej
 Dwustopniowa 100
 opcje modelu 80
 opcje modelu KNN 90, 91
 opcje modelu szeregów czasowych 99
 Opcje uogólnionego modelu
 liniowego 85
 Opcje zmiennej Dwustopniowa 100
 Opcje zmiennej grupowania
 dzielącego 83
 Opcje zmiennej K-średnie 92
 Opcje zmiennej PCA 101
 Opcje zmiennej sieci Bayesa 93
 opcje zmiennych 79
 Opcje zmiennych drzewa decyzyjnego 88
 Opcje zmiennych szeregu czasowego 96
 PCA 101
 regresja liniowa 90
 Sieć Bayesa 93
 Szereg czasowy 94
 Uogólnione liniowe 84
 Użytkowy uogólniony model liniowy 84,
 110
 Wartościowa informacja z modelu
 K-średnie 104
 zarządzanie modelami 102
 IBM SPSS Modeler 1
 dokumentacja 3
 eksploracja bazy danych 7
 IBM SPSS Modeler Server 1
 IBM SPSS Modeler Solution Publisher
 Modele Oracle Data Mining 29
 InfoSphere Warehouse (IBM), patrz ISW 50
 InfoSphere Warehouse Data Mining
 drzewa decyzyjne 57
 modele użytkowe 70
 modelowanie asocjacyjne 58
 przykładowe strumienie 71
 systematyka 60
 węzeł regresji 63
 węzeł Sekwencje 61
 interpolacja wartości, szereg czasowy IBM
 Netezza Analytics 95
 ISW
 integracja z produktem IBM SPSS
 Modeler 50
 Karta Serwer 55
 Połączenie ODBC 50

K

K-średnie
 IBM DB2 for z/OS 117, 124
 IBM Netezza Analytics 92, 104
 Oracle Data Mining 39
 kara za złożoność 17, 18, 19
 Karta Serwer
 ISW 55
 kategoryzacja danych
 Modele Oracle 46
 klucz
 klucze modeli 9
 konfigurowanie
 IBM DB2 for z/OS i IBM Analytics
 Accelerator for z/OS 114
 koszty
 Oracle 30
 koszty błędnej klasyfikacji
 Oracle 30
 kryterium podziału
 Oracle K-średnie 39
L
 liczba skupień
 Oracle K-średnie 39
 Oracle O-Cluster 38
 liniowy algorytm domyślny
 Oracle Support Vector Machine 33
 liść, w modelach drzewa Netezza 87, 118

M

MDL 32
 metoda normalizacji
 Oracle K-średnie 39
 Oracle NMF 40
 Oracle Support Vector Machine 33
 metryka zanieczyszczenia
 Oracle Apriori 37
 miara zanieczyszczenia entropii 88
 miara zanieczyszczenia Gini 88
 miary zanieczyszczenia
 drzewo decyzyjne 119
 Netezza Decision Tree 88
 Microsoft
 Grupowanie sekwencyjne 11
 modelowanie drzewa decyzyjnego 11,
 13, 21
 modelowanie Naive Bayes 13, 21
 Modelowanie Naive Bayes 11
 modelowanie regresji liniowej 13, 21
 modelowanie regresji logistycznej 13, 21
 modelowanie reguł asocjacyjnych 11, 13,
 21
 modelowanie sieci neuronowej 13, 21
 modelowanie skupień 11, 13, 21
 regresja liniowa 11
 Regresja logistyczna 11
 Sieć neuronowa 11
 Usługi Analysis Services 11, 13, 21
 zarządzanie modelami 15
 Microsoft Analysis Services 22, 23, 24
 min.-maks.
 normalizowanie danych 33, 46
 Minimum Description Length (MDL)
 Oracle Data Mining 42, 43

Minimum Description Length (MDL)
 Oracle Data Mining 42, 43
 modele
 budowanie modeli w bazie danych 8
 eksportowanie 9
 lista DB2 54
 lista Netezza 80
 ocena 25, 48, 72
 ocenianie modeli w bazie danych 8
 problemy ze spójnością 9
 przeglądanie w DB2 55
 przeglądanie w Oracle 32
 zapisywanie 9
 zarządzanie DB2 54
 zarządzanie Netezza 80
 zarządzanie usługami Analysis
 Services 15
 modele ARIMA
 IBM Netezza Analytics 94, 98
 modele drzew decyzyjnych
 InfoSphere Warehouse Data Mining 57
 Modele KNN
 IBM Netezza Analytics 106
 modele Naive Bayes
 IBM Netezza Analytics 106
 sieć Oracle Adaptive Bayes 32
 modele najbliższego sąsiedztwa
 IBM Netezza Analytics 90, 91, 106
 modele PCA
 IBM Netezza Analytics 101, 107, 108
 modele przycięte Naive Bayes
 sieć Oracle Adaptive Bayes 32
 modele reguł asocjacyjnych
 Microsoft 18
 modele sieci bayesowskiej
 IBM Netezza Analytics 93, 94, 104, 105
 modele użytkowe
 IBM DB2 for z/OS 123, 124, 125
 IBM Netezza Analytics 84, 103, 104,
 105, 106, 107, 108, 109, 110, 111
 InfoSphere Warehouse Data Mining 70
 modele z jednym predyktorem
 Sieć Oracle Adaptive Bayes 32
 modele z wieloma predyktorami
 sieć Oracle Adaptive Bayes 32
 modelowanie asocjacyjne
 InfoSphere Warehouse Data Mining 58
 modelowanie w bazie danych 22
 IBM Netezza Analytics 75, 76, 78, 79
 Oracle 27, 28, 29, 30
 modelowanie w programie DB2 for z/OS
 IBM DB2 for z/OS 113, 115, 116

N

naive bayes
 ocenianie — opcje podsumowania 22
 ocenianie — opcje serwera 21
 opcje modelu 17
 opcje serwera 17
 opcje zaawansowane 18
 Naive Bayes
 IBM DB2 for z/OS 118, 124
 IBM Netezza Analytics 93, 105
 InfoSphere Warehouse Data Mining 68
 Oracle Data Mining 31

nazwa hosta
 Połączenie z bazą Oracle 28

Netezza
 zarządzanie modelami 80

NMF
 Oracle Data Mining 40

normalizowanie danych
 Modele Oracle 46

O

O-Cluster
 Oracle Data Mining 38, 39

ocena 25, 48, 72

ocenywanie 8, 102, 123

ocenywanie modelu
 InfoSphere Warehouse Data Mining 53

ODBC
 konfigurowanie 13
 konfigurowanie na potrzeby IBM DB2 for z/OS 116
 konfigurowanie na potrzeby IBM Netezza Analytics 75, 76, 78, 79
 konfigurowanie na potrzeby Oracle 27, 28, 29, 30
 konfigurowanie produktu ISW 50
 konfigurowanie programu SQL Server 13

odchylenie standardowe
 Oracle Support Vector Machine 34

ODM. Patrz Oracle Data Mining 27

opcje budowania
 IBM DB2 for z/OS 117, 119, 120, 121, 122
 IBM Netezza Analytics 81, 82, 83, 88, 89, 90, 92, 94, 97, 98, 100

opcje modelu
 IBM DB2 for z/OS 116
 IBM Netezza Analytics 80, 85, 90, 91, 99

opcje zaawansowane
 ISW Data Mining 56

opcje zmiennych
 IBM DB2 for z/OS 116, 117, 118, 122
 IBM Netezza Analytics 79, 83, 88, 92, 93, 96, 100, 101

węzły modelowania 58

Oracle Data Miner 46

Oracle Data Mining 27
 AI (ang. Attribute Importance) 43, 44
 Apriori 41, 42
 drzewo decyzyjne 37, 38
 K-średnie 39
 konfigurowanie za pomocą produktu IBM SPSS Modeler 27, 28, 29, 30
 koszty błędnej klasyfikacji 45
 Minimum Description Length (MDL) 42, 43
 Naive Bayes 31
 NMF 40
 O-Cluster 38, 39
 przygotowanie danych 46
 przykłady 47, 48
 sieć Adaptive Bayes 32, 33
 sprawdzanie spójności 44
 Support Vector Machine 33, 34
 Uogólnione modele liniowe (GLM) 35, 36
 zarządzanie modelami 44, 45

P

plik tnsnames.ora 28

pojedyncza wartość graniczna
 Oracle Naive Bayes 31

port
 Połączenie z bazą Oracle 28

prawdopodobieństwa a priori
 Oracle Data Mining 35

przykłady
 Applications — podręcznik 3
 eksploracja bazy danych 24, 25, 26, 48, 71, 72, 73
 przegląd 4
 przykłady aplikacji 3

R

regresja liniowa
 IBM DB2 for z/OS 120
 IBM Netezza Analytics 81, 90, 109
 ocenywanie — opcje podsumowania 22
 ocenywanie — opcje serwera 21
 opcje modelu 17
 opcje serwera 17
 opcje zaawansowane 18

regresja logistyczna
 ocenywanie — opcje podsumowania 22
 ocenywanie — opcje serwera 21
 opcje modelu 17
 opcje serwera 17
 opcje zaawansowane 18

reguły asocjacyjne
 ocenywanie — opcje podsumowania 22
 ocenywanie — opcje serwera 21
 opcje modelu 17
 opcje serwera 17
 opcje zaawansowane 18

S

serwer
 uruchamianie usług Analysis Services 17, 21, 22

SID
 Połączenie z bazą Oracle 28

sieć Adaptive Bayes
 Oracle Data Mining 32, 33

sieć neuronowa
 ocenywanie — opcje podsumowania 22
 ocenywanie — opcje serwera 21
 opcje modelu 17
 opcje serwera 17
 opcje zaawansowane 18

Solution Publisher
 Modele Oracle Data Mining 29

SQL Server 17, 21, 22
 konfigurowanie 13
 Połączenie ODBC 13

statystyki z
 normalizowanie danych 33, 46

strumienie
 Przykłady dotyczące InfoSphere Warehouse Data Mining 71

Support Vector Machine
 Oracle Data Mining 33, 34

SVM. Patrz Support Vector Machine 33

systematyka
 InfoSphere Warehouse Data Mining 60

Szereg czasowy
 IBM Netezza Analytics 96, 97, 98, 99
 InfoSphere Warehouse Data Mining 68, 69, 70

szereg czasowy (IBM Netezza Analytics) 109, 110

Szereg czasowy (IBM Netezza Analytics) 94

szereg czasowy (Microsoft) 19
 opcje modelu 19
 opcje ustawień 20
 opcje zaawansowane 19

T

tolerancja zbieżności
 Oracle Support Vector Machine 34

tryb audytu danych 25, 48, 72

U

uogólnione modele liniowe
 IBM Netezza Analytics 84, 85, 86, 110

Uogólnione modele liniowe (GLM)
 Oracle Data Mining 35, 36

Usługi Analysis Services
 Drzewa decyzyjne 24
 przykłady 24
 zarządzanie modelami 15

W

waga instancji, w modelach drzewa
 Netezza 87

waga klasy, w modelach drzewa Netezza 87

walidacja krzyżowa
 Oracle Naive Bayes 31

wartość graniczna parami
 Oracle Naive Bayes 31

wdrażanie 26, 48, 73

Węzeł grupowania
 InfoSphere Warehouse Data Mining 65

Węzeł publikacji
 Modele Oracle Data Mining 29

węzeł regresji
 InfoSphere Warehouse Data Mining 63

węzeł regresji logistycznej
 InfoSphere Warehouse Data Mining 68

węzeł Sekwencji
 InfoSphere Warehouse Data Mining 61

węzły
 generowanie 24

węzły modelowania
 Drzewa decyzyjne Microsoft 15
 modelowanie w bazie danych 8, 11, 13, 15, 21
 modelowanie w bazie danych da ISW 50

MS Association Rules 15

MS Clustering 15

MS Linear Regression 15

MS Logistic Regression 15

MS Naive Bayes 15

MS Neural Network 15

MS Sequence Clustering 15

MS Time Series 15

wyglądanie wykładnicze
 IBM Netezza Analytics 94
wymogi
 IBM DB2 for z/OS 113

Z

zmienna unikalna
 Oracle Apriori 37, 42
 Oracle Data Mining 29
 Oracle K-średnie 39
 Oracle MDL 43
 Oracle Naive Bayes 31
 Oracle NMF 40
 Oracle O-Cluster 38
 Oracle Support Vector Machine 33
 sieć Oracle Adaptive Bayes 32
zmiennie dzielące na podzbiory
 wybór 41



Drukowane w USA