

Guide d'exploration de base de données  
de IBM SPSS Modeler 15



*Remarque* : Avant d'utiliser ces informations et le produit qu'elles concernent, lisez les informations générales sous Remarques sur p. .

Cette version s'applique à IBM SPSS Modeler 15 et à toutes les publications et modifications ultérieures jusqu'à mention contraire dans les nouvelles versions.

Les captures d'écran des produits Adobe sont reproduites avec l'autorisation de Adobe Systems Incorporated.

Les captures d'écran des produits Microsoft sont reproduites avec l'autorisation de Microsoft Corporation.

Matériel sous licence - Propriété d'IBM

© **Copyright IBM Corporation 1994, 2012.**

Droits limités pour les utilisateurs au sein d'administrations américaines : utilisation, copie ou divulgation soumise au GSA ADP Schedule Contract avec IBM Corp.

---

# Préface

IBM® SPSS® Modeler est le puissant utilitaire de Data mining de IBM Corp.. SPSS Modeler aide les entreprises et les organismes à améliorer leurs relations avec les clients et les citoyens grâce à une compréhension approfondie des données. A l'aide des connaissances plus précises obtenues par le biais de SPSS Modeler, les entreprises et les organismes peuvent conserver les clients rentables, identifier les opportunités de vente croisée, attirer de nouveaux clients, détecter les éventuelles fraudes, réduire les risques et améliorer les services gouvernementaux.

L'interface visuelle de SPSS Modeler met à contribution les compétences professionnelles de l'utilisateur, ce qui permet d'obtenir des modèles prédictifs plus efficaces et de trouver des solutions plus rapidement. SPSS Modeler dispose de nombreuses techniques de modélisation, telles que les algorithmes de prévision, de classification, de segmentation et de détection d'association. Une fois les modèles créés, l'utilisateur peut utiliser IBM® SPSS® Modeler Solution Publisher pour les remettre aux responsables, où qu'ils se trouvent dans l'entreprise, ou pour les transférer vers une base de données.

## ***A propos de IBM Business Analytics***

Le logiciel IBM Business Analytics fournit des informations complètes, cohérentes et précises que les preneurs de décision utilisent avec confiance pour améliorer la performance du marché. Un portefeuille étendu d'outils de [business intelligence](#), d'[analyses prédictives](#), de [performance financière et de gestion de stratégie](#), et des [applications analytiques](#) offre des connaissances claires, immédiates et applicables pour améliorer l'efficacité actuelle ainsi que la capacité de prévoir les résultats futurs. Combinées avec de riches solutions industrielles, des pratiques éprouvées et des services professionnels, les organisations de toutes tailles peuvent atteindre la productivité la plus élevée, automatiser des décisions en toute tranquillité et fournir de meilleurs résultats.

Dans le cadre de ce portefeuille, le logiciel IBM SPSS Predictive Analytics aide les organisations à prévoir des événements futurs et à agir en conséquence pour mener à de meilleurs résultats. Des clients dans le domaine commercial, gouvernemental et académique à travers le monde font confiance à la technologie IBM SPSS et considèrent qu'elle représente un avantage compétitif pour attirer, retenir et ajouter des clients, tout en réduisant la fraude et en atténuant les risques. En incorporant le logiciel IBM SPSS dans leur opérations quotidiennes, les organisations deviennent des entreprises prédictives – capables de diriger et d'automatiser les décisions pour atteindre les buts qu'ils se sont fixés et obtenir des avantages compétitifs sensibles. Pour informations supplémentaires ou pour joindre un revendeur, visitez le site <http://www.ibm.com/spss>.

## ***Assistance technique***

L'assistance technique est à la disposition des clients pour la maintenance des produits. Les clients peuvent contacter l'assistance technique pour obtenir de l'aide concernant l'utilisation des produits IBM Corp. ou l'installation dans l'un des environnements matériels pris en charge. Pour joindre l'assistance technique, consultez le site Web de IBM Corp. à l'adresse <http://www.ibm.com/support>. Lorsque vous contactez l'assistance technique, n'oubliez pas de préparer vos identifiants, le nom de votre société et votre contrat d'assistance.

---

# Contenu

## **1 A propos de IBM SPSS Modeler 1**

À propos de IBM SPSS Modeler . . . . .	1
Produits IBM SPSS Modeler . . . . .	1
IBM SPSS Modeler . . . . .	1
IBM SPSS Modeler Server . . . . .	2
IBM SPSS Modeler Administration Console . . . . .	2
IBM SPSS Modeler Batch . . . . .	2
IBM SPSS Modeler Solution Publisher . . . . .	3
IBM SPSS Modeler Server Adaptateurs pour IBM SPSS Collaboration and Deployment Services . . . . .	3
Éditions de IBM SPSS Modeler . . . . .	3
Documentation de IBM SPSS Modeler . . . . .	4
Documentation de SPSS Modeler Professional . . . . .	4
Documentation de SPSS Modeler Premium . . . . .	5
Exemples d'application . . . . .	6
Dossier Demos . . . . .	6

## **2 Exploration des bases de données 8**

Modélisation de base de données - Présentation . . . . .	8
Composants requis . . . . .	9
Construction du modèle . . . . .	10
Préparation des données . . . . .	11
Scoring de modèles . . . . .	11
Exportation et enregistrement de modèles de base de données . . . . .	12
Cohérence d'un modèle . . . . .	13
Affichage et exportation de code SQL généré . . . . .	13

## **3 Modélisation de base de données à l'aide de Microsoft Analysis Services 14**

IBM SPSS Modeler et Microsoft Analysis Services . . . . .	14
Conditions requises pour l'intégration à Microsoft Analysis Services . . . . .	15
Activation de l'intégration à Analysis Services . . . . .	17
Création de modèles à l'aide d'Analysis Services . . . . .	20
Gestion des modèles Analysis Services . . . . .	21
Paramètres communs à tous les noeuds d'algorithme . . . . .	22

Arbre de décision MS - Options expert . . . . .	25
Classification MS - Options expert . . . . .	26
Options expert MS Naive Bayes. . . . .	27
Régression linéaire MS - Options expert . . . . .	28
Réseau de neurones MS - Options expert . . . . .	29
Régression logistique MS - Options expert. . . . .	30
Noeud Règles d'association MS . . . . .	30
Noeud Séries temporelles MS . . . . .	32
Noeud Classification de séquences MS. . . . .	36
Scoring des modèles Analysis Services . . . . .	38
Paramètres communs à tous les modèles Analysis Services . . . . .	39
Nugget du modèle Séries temporelles MS . . . . .	42
Nugget de modèle de classification de séquences MS . . . . .	46
Exportation de modèles et génération de noeuds. . . . .	46
Exemples d'exploration Analysis Services . . . . .	46
Exemples de flux : Arbres décision . . . . .	47

## **4 Modélisation de bases de données à l'aide d'Oracle Data Mining** **55**

A propos d'Oracle Data Mining . . . . .	55
Conditions requises pour l'intégration à Oracle. . . . .	55
Activation de l'intégration à Oracle. . . . .	56
Création de modèles à l'aide d'Oracle Data Mining . . . . .	59
Options du serveur de modèles Oracle. . . . .	60
Coûts de mauvaise réaffectation . . . . .	61
Oracle Naive Bayes . . . . .	62
Options des modèles Naive Bayes . . . . .	63
Options expert Naive Bayes. . . . .	64
Oracle Adaptive Bayes . . . . .	64
Options des modèles Adaptive Bayes . . . . .	65
Options expert Adaptive Bayes . . . . .	67
Oracle Support Vector Machine (SVM) . . . . .	67
Options des modèles Oracle SVM . . . . .	68
Options expert Oracle SVM . . . . .	69
Options de pondérations Oracle SVM . . . . .	71
Modèles linéaires généralisés (MLG) Oracle. . . . .	72
Options des modèles Oracle MLG . . . . .	72
Options expert Oracle MLG . . . . .	73
Options de pondération Oracle MLG . . . . .	74

Arbre décision Oracle . . . . .	75
Options du modèle Arbre décision . . . . .	76
Arbre de décision - Options expert . . . . .	77
O-Cluster Oracle . . . . .	78
Options du modèle O-Cluster . . . . .	78
Options expert O-Cluster . . . . .	79
Algorithme Oracle k-Means . . . . .	79
Options du modèle k-Means . . . . .	80
Options expert de k-Means . . . . .	81
Oracle Nonnegative Matrix Factorization (NMF) . . . . .	82
Options du modèle NMF . . . . .	82
Options expert NMF . . . . .	83
Apriori Oracle . . . . .	84
Options des champs Apriori . . . . .	84
Options du modèle Apriori . . . . .	87
Description de longueur minimale d'Oracle (MDL) . . . . .	88
Options du modèle MDL . . . . .	89
Importance de l'attribut d'Oracle (IA) . . . . .	90
IA - Options du modèle . . . . .	90
IA - Options de sélection . . . . .	91
Nugget du modèle IA - Onglet Modèle . . . . .	92
Gestion des modèles Oracle . . . . .	93
Nugget de modèle Oracle - Onglet Serveur . . . . .	93
Nugget de modèle Oracle - Onglet Récapitulatif . . . . .	94
Nugget de modèle Oracle - Onglet Paramètres . . . . .	95
Liste des modèles Oracle . . . . .	95
Oracle Data Miner . . . . .	97
Préparation des données . . . . .	99
Oracle Data Mining - Exemples . . . . .	99
Exemple de flux : Envoi de données . . . . .	100
Exemple de flux : Exploration des données . . . . .	101
Exemple de flux : Créer un modèle . . . . .	102
Exemple de flux : Evaluation d'un modèle . . . . .	103
Exemple de flux : Déploiement d'un modèle . . . . .	106

## **5 Modélisation de bases de données à l'aide d'IBM InfoSphere Warehouse 107**

IBM InfoSphere Warehouse et IBM SPSS Modeler . . . . .	107
Conditions requises pour l'intégration à IBM InfoSphere Warehouse . . . . .	107
Activation de l'intégration à IBM InfoSphere Warehouse . . . . .	108

Création de modèles à l'aide de IBM InfoSphere Warehouse Data Mining . . . . .	115
Détermination de scores et déploiement du modèle . . . . .	116
Gestion des modèles DB2 . . . . .	117
Liste des modèles de base de données . . . . .	118
Navigation dans les modèles . . . . .	119
Exportation de modèles et génération de noeuds . . . . .	119
Paramètres de noeud communs à tous les algorithmes . . . . .	119
Arbre décision ISW . . . . .	123
Arbre décision ISW - Options du modèle . . . . .	124
Arbre de décision ISW - Options expert . . . . .	125
Association ISW . . . . .	125
Association ISW - Options de champs . . . . .	126
Association ISW - Options de modèle . . . . .	129
Association ISW - Options expert . . . . .	130
Options de taxonomie d'ISW . . . . .	131
Séquence ISW . . . . .	134
Séquence ISW - Options de modèle . . . . .	135
Séquence ISW - Options expert . . . . .	136
Régression ISW . . . . .	137
Régression ISW - Options de modèle . . . . .	139
Régression ISW - Options expert . . . . .	140
Classification ISW . . . . .	142
Classification ISW - Options de modèle . . . . .	143
Options expert de la classification ISW . . . . .	145
Naive Bayes ISW . . . . .	148
Options des modèles Naive Bayes d'ISW . . . . .	148
Régression logistique ISW . . . . .	149
Régression logistique d'ISW - Options de modèle . . . . .	149
Séries temporelles ISW . . . . .	150
Options des champs des séries temporelles ISW . . . . .	151
Options du modèle de séries temporelles ISW . . . . .	152
Options de l'expert de séries temporelles ISW . . . . .	153
Affichage des modèles de séries temporelles ISW . . . . .	154
Nuggets de modèle ISW Data Mining . . . . .	155
Nugget de modèle ISW - Onglet Serveur . . . . .	155
Nugget de modèle ISW - Onglet Paramètres . . . . .	156
Nugget de modèle ISW - Onglet Récapitulatif . . . . .	157
Exemples d'ISW Data mining . . . . .	158
Exemple de flux : Envoi de données . . . . .	158
Exemple de flux : Exploration des données . . . . .	159
Exemple de flux : Créer un modèle . . . . .	160

Exemple de flux : Evaluation d'un modèle . . . . .	161
Exemple de flux : Déploiement d'un modèle . . . . .	163

## **6 Modélisation de la base de données avec IBM Netezza Analytics 165**

IBM SPSS Modeler et IBM Netezza Analytics . . . . .	165
Conditions requises pour l'intégration à IBM Netezza Analytics . . . . .	165
Activation de l'intégration à IBM Netezza Analytics . . . . .	166
Configuration de IBM Netezza Analytics . . . . .	166
Création d'une source de données ODBC pour IBM Netezza Analytics . . . . .	166
Activation de l'intégration de IBM Netezza Analytics dans IBM SPSS Modeler . . . . .	168
Activation de la génération SQL et de l'optimisation . . . . .	169
Création de modèles avec IBM Netezza Analytics . . . . .	170
Modèles Netezza - Options de champs . . . . .	171
Modèles Netezza - Options du serveur . . . . .	172
Modèles Netezza - Options du modèle . . . . .	174
Arbres décision Netezza . . . . .	175
Pondérations d'instance et pondérations de classe . . . . .	176
Options du champ Arbre décision Netezza . . . . .	177
Options de création d'arbre décision Netezza . . . . .	178
Netezza K-means . . . . .	183
Options du champ K-Means Netezza . . . . .	183
Options de création K-Means Netezza . . . . .	185
Bayes Net Netezza . . . . .	186
Options de champs du réseau Bayes Net Netezza . . . . .	186
Options de création du réseau Bayes Net Netezza . . . . .	187
Naive Bayes Netezza . . . . .	188
KNN Netezza . . . . .	189
Options de modèle KNN Netezza - Général . . . . .	189
Options de modèle KNN Netezza - Options de scoring . . . . .	191
Classification par division Netezza . . . . .	193
Options de champ de classification par division Netezza . . . . .	193
Options de création de classification par division Netezza . . . . .	195
ACP Netezza . . . . .	196
Options de champs ACP Netezza . . . . .	196
Options de création ACP Netezza . . . . .	198
Arbre de régression Netezza . . . . .	199
Options de création d'arbre de régression Netezza - Développement de l'arbre . . . . .	199
Options de création d'arbre de régression Netezza - Élagage de l'arbre . . . . .	201



Régression linéaire Netezza . . . . .	202
Options de création de régression linéaire Netezza . . . . .	202
Séries temporelles Netezza . . . . .	204
Interpolation des valeurs dans les séries temporelles Netezza . . . . .	205
Options des champs de séries temporelles Netezza . . . . .	206
Options de création de séries temporelles Netezza . . . . .	208
Options du modèle de série temporelle Netezza . . . . .	213
Linéaire généralisé Netezza . . . . .	215
Options de modèle linéaire généralisé Netezza - Généralités . . . . .	215
Options de modèle linéaire généralisé Netezza - Interactions . . . . .	217
Options de modèle linéaire généralisé Netezza - Options de scoring . . . . .	220
Gestion des modèles IBM Netezza Analytics . . . . .	220
Détermination de scores des modèles IBM Netezza Analytics . . . . .	220
Onglet serveur du nugget de modèle Netezza . . . . .	221
Nuggets de modèle Arbre de décision Netezza . . . . .	222
Nugget du modèle K-Means Netezza . . . . .	224
Nuggets de modèle Bayes Net Netezza . . . . .	225
Nuggets de modèle Naive Bayes Netezza . . . . .	227
Nuggets de modèles KNN Netezza . . . . .	228
Nuggets des modèles de classification par division Netezza . . . . .	230
Nuggets de modèles ACP Netezza . . . . .	231
Nuggets de modèle Arbre de régression Netezza . . . . .	232
Nuggets du modèle de régression linéaire Netezza . . . . .	234
Nugget du modèle Séries temporelles Netezza . . . . .	234
Nugget de modèle linéaire généralisé Netezza . . . . .	235

## ***Annexe***

<b><i>A Remarques</i></b>	<b>238</b>
---------------------------	------------

<b><i>Index</i></b>	<b>241</b>
---------------------	------------



# *A propos de IBM SPSS Modeler*

## *À propos de IBM SPSS Modeler*

IBM® SPSS® Modeler est un ensemble d'outils de data mining qui vous permet de développer rapidement, grâce à vos compétences professionnelles, des modèles prédictifs et de les déployer dans des applications professionnelles afin de faciliter la prise de décision. Conçu autour d'un modèle confirmé, le modèle CRISP-DM, SPSS Modeler prend en charge l'intégralité du processus de Data mining, des données à l'obtention de meilleurs résultats commerciaux.

SPSS Modeler propose différentes méthodes de modélisation issues des domaines de l'apprentissage automatique, de l'intelligence artificielle et des statistiques. Les méthodes disponibles dans la palette Modélisation vous permettent d'extraire de nouvelles informations de vos données et de développer des modèles prédictifs. Chaque méthode possède ses propres avantages et est donc plus adaptée à certains types de problème spécifiques.

Il est possible d'acquérir SPSS Modeler comme produit autonome ou de l'utiliser en tant que client en combinaison avec SPSS Modeler Server. Plusieurs autres options sont également disponibles, telles que décrites dans les sections suivantes. Pour plus d'informations, consultez <http://www.ibm.com/software/analytics/spss/products/modeler/>.

## *Produits IBM SPSS Modeler*

La famille des produits IBM® SPSS® Modeler et les logiciels associés sont composés des éléments suivants.

- 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 adaptateurs pour IBM SPSS Collaboration and Deployment Services

## *IBM SPSS Modeler*

SPSS Modeler est une version complète du produit que vous installez et exécutez sur votre ordinateur personnel. Pour obtenir de meilleures performances lors du traitement d'ensembles de données volumineux, vous pouvez exécuter SPSS Modeler en mode local, comme produit autonome, ou l'utiliser en mode réparti, en association avec IBM® SPSS® Modeler Server.

Avec SPSS Modeler, vous pouvez créer des modèles prédictifs précis rapidement et de manière intuitive, sans aucune programmation. L'interface visuelle unique vous permet de visualiser facilement le processus de Data mining. Grâce aux analyses avancées intégrées au produit, vous pouvez découvrir des motifs et tendances masqués dans vos données. Vous pouvez modéliser les résultats et comprendre les facteurs qui les influencent, afin d'exploiter les opportunités commerciales et de réduire les risques.

SPSS Modeler est disponible en deux éditions : SPSS Modeler Professional et SPSS Modeler Premium. [Pour plus d'informations, reportez-vous à la section Éditions de IBM SPSS Modeler dans \*Guide de l'utilisateur de IBM SPSS Modeler 15\*.](#)

### ***IBM SPSS Modeler Server***

Grâce à une architecture client/serveur, SPSS Modeler adresse les demandes d'opérations très consommatrices de ressources à un logiciel serveur puissant. Il offre ainsi des performances accrues sur des ensembles de données plus volumineux.

SPSS Modeler Server est un produit avec licence distincte qui s'exécute en permanence en mode d'analyse réparti sur un hôte de serveur en combinaison avec une ou plusieurs installations de IBM® SPSS® Modeler. Ainsi, SPSS Modeler Server fournit des performances supérieures sur de grands ensembles de données car les opérations nécessitant beaucoup de mémoire peuvent être effectuées sur le serveur sans télécharger de données sur l'ordinateur client. IBM® SPSS® Modeler Server prend également en charge l'optimisation SQL et propose des fonctionnalités de modélisation dans la base de données pour des performances et une automatisation améliorées.

### ***IBM SPSS Modeler Administration Console***

Le Modeler Administration Console est une application graphique permettant de gérer de nombreuses options de SPSS Modeler Server qui peuvent également être configurées au moyen d'un fichier d'options. Cette application offre une interface utilisateur sous forme de console permettant de surveiller et de configurer les installations SPSS Modeler Server ; elle est disponible gratuitement pour les clients actuels de SPSS Modeler Server. L'application ne peut être installée que sur des ordinateurs Windows ; en revanche, elle peut administrer un serveur installé sur n'importe quelle plate-forme prise en charge.

### ***IBM SPSS Modeler Batch***

Alors que le Data mining est généralement un processus interactif, il est également possible d'exécuter SPSS Modeler à partir d'une ligne de commande sans recourir à l'interface utilisateur graphique. Par exemple, vous pouvez avoir des tâches longue durée ou répétitives à exécuter sans intervention de l'utilisateur. SPSS Modeler Batch est une version spécifique du produit qui prend en charge toutes les fonctions d'analyse de SPSS Modeler sans avoir besoin d'accéder à l'interface utilisateur standard. Une licence SPSS Modeler Server est nécessaire pour utiliser SPSS Modeler Batch.

## **IBM SPSS Modeler Solution Publisher**

SPSS Modeler Solution Publisher est un outil qui permet de créer une version « packagée » d'un flux SPSS Modeler qui peut être exécutée par un moteur Runtime externe ou intégrée dans une application externe. Ainsi, vous pouvez publier et déployer des flux SPSS Modeler complets dans des environnements où SPSS Modeler n'est pas installé. SPSS Modeler Solution Publisher est fourni avec le service IBM SPSS Collaboration and Deployment Services - Scoring et nécessite une licence distincte. Avec cette licence, vous recevez SPSS Modeler Solution Publisher Runtime qui vous permet d'exécuter les flux publiés.

## **IBM SPSS Modeler Server Adaptateurs pour IBM SPSS Collaboration and Deployment Services**

Différents adaptateurs pour IBM® SPSS® Collaboration and Deployment Services sont disponibles et permettent à SPSS Modeler et SPSS Modeler Server d'interagir avec un référentiel IBM SPSS Collaboration and Deployment Services. Ainsi, un flux SPSS Modeler déployé sur le référentiel peut être partagé par différents utilisateurs ou peut être accessible depuis l'application client léger IBM SPSS Modeler Advantage. Installez l'adaptateur sur le système qui héberge le référentiel.

## **Éditions de IBM SPSS Modeler**

SPSS Modeler est disponible dans les éditions suivantes.

### **SPSS Modeler Professional**

SPSS Modeler Professional offre tous les outils nécessaires à l'utilisation de la plupart des types de données structurées, tels que les comportements et interactions suivis dans les systèmes CRM, les caractéristiques sociodémographiques, les comportements d'achat et les données de vente.

### **SPSS Modeler Premium**

SPSS Modeler Premium est un produit avec licence distincte qui étend le champ d'applications de SPSS Modeler Professional afin de pouvoir traiter des données spécialisées telles que celles utilisées pour les analyses d'entités ou les réseaux sociaux ainsi que des données de texte non structurées. SPSS Modeler Premium comprend les composants suivants :

**IBM® SPSS® Modeler Entity Analytics** ajoute une dimension entièrement nouvelle aux analyses prédictives IBM® SPSS® Modeler. Alors que les analyses prédictives essaient de prévoir les comportements futurs à partir de données passées, les analyses d'entités se concentrent sur l'amélioration de la cohérence des données actuelles en résolvant les conflits d'identités dans les enregistrements eux-mêmes. Une identité peut être celle d'un individu, d'une organisation, d'un objet ou d'une autre entité pour laquelle une ambiguïté peut exister. La résolution d'identité peut être vitale dans de nombreux domaines, y compris la gestion de la relation client, la détection de la fraude, le blanchiment d'argent et la sécurité nationale et internationale.

**IBM SPSS Modeler Social Network Analysis** transforme les informations sur les relations en champs qui caractérisent le comportement social des individus et des groupes. Grâce aux données qui décrivent les relations qui sous-tendent les réseaux sociaux, IBM® SPSS® Modeler Social Network Analysis identifie les chefs sociaux qui influencent le comportement des autres individus du réseau. De plus, il est possible de déterminer les individus qui sont le plus influencés par les autres participants du réseau. En combinant ces résultats avec d'autres mesures, il est possible de créer des profils détaillés des individus sur lesquels baser vos modèles prédictifs. Les modèles qui contiennent ces informations sociales seront plus efficaces que les modèles qui en sont dépourvus.

**Text Analytics for IBM® SPSS® Modeler** utilise des technologies linguistiques avancées et le traitement du langage naturel pour traiter rapidement une large variété de données textuelles non structurées, en extraire les concepts clés et les organiser pour les regrouper dans des catégories. Les concepts extraits et les catégories peuvent ensuite être combinés aux données structurées existantes, telles que les données démographiques, et appliqués à la modélisation grâce à la gamme complète d'outils de Data mining de SPSS Modeler, afin de favoriser une prise de décision précise et efficace.

## ***Documentation de IBM SPSS Modeler***

Une documentation au format d'aide en ligne est disponible dans le menu Aide de SPSS Modeler. Vous y trouverez la documentation de SPSS Modeler, SPSS Modeler Server et de SPSS Modeler Solution Publisher, ainsi que le Guide des applications et d'autres documentations utiles.

La documentation complète de chaque produit (y compris les instructions d'installation) au format PDF est disponible dans le dossier *Documentation* de chaque DVD de produit. Ces documents d'installation peuvent également être téléchargés sur Internet à l'adresse <http://www-01.ibm.com/support/docview.wss?uid=swg27023172>.

La documentation dans les deux formats est également disponible depuis le Centre d'informations SPSS Modeler à l'adresse <http://publib.boulder.ibm.com/infocenter/spssmodl/v15r0m0/>.

## ***Documentation de SPSS Modeler Professional***

La suite de documentation SPSS Modeler Professional (à l'exception des instructions d'installation) est la suivante.

- **Guide de l'utilisateur IBM SPSS Modeler.** Introduction générale à SPSS Modeler : création de flux de données, traitement des valeurs manquantes, création d'expressions CLEM, utilisation des projets et des rapports et regroupement des flux pour le déploiement dans IBM SPSS Collaboration and Deployment Services, des applications prédictives ou IBM SPSS Modeler Advantage.
- **Noeuds de Source, d'exécution et de sortie IBM SPSS Modeler.** Descriptions de tous les noeuds utilisés pour lire, traiter et renvoyer les données de sortie dans différents formats. En pratique, cela signifie tous les noeuds autres que les noeuds de modélisation.
- **IBM SPSS Modeler Noeuds de modélisation.** Description de tous les noeuds utilisés pour créer des modèles de Data mining. IBM® SPSS® Modeler propose différentes méthodes de modélisation issues des domaines de l'apprentissage automatique, de l'intelligence artificielle

et des statistiques. [Pour plus d'informations, reportez-vous à la section Description des noeuds de modélisation dans le chapitre 3 dans \*Noeuds de modélisation de IBM SPSS Modeler 15\*.](#)

- **Guide des Algorithmes IBM SPSS Modeler.** Descriptions des fondements mathématiques des méthodes de modélisation utilisées dans SPSS Modeler. Ce guide est disponible au format PDF uniquement.
- **Guide des applications IBM SPSS Modeler.** Les exemples de ce guide fournissent des introductions brèves et ciblées aux méthodes et techniques de modélisation. Une version en ligne de ce guide est également disponible dans le menu Aide. [Pour plus d'informations, reportez-vous à la section Exemples d'application dans \*Guide de l'utilisateur de IBM SPSS Modeler 15\*.](#)
- **Génération de scripts et automatisation IBM SPSS Modeler.** Informations sur l'automatisation du système via la génération de scripts, y compris les propriétés permettant de manipuler les noeuds et les flux.
- **IBM SPSS Modeler Guide de déploiement.** Informations sur l'exécution des scénarios et des flux SPSS Modeler comme étapes des tâches d'exécution sous IBM® SPSS® Collaboration and Deployment Services Deployment Manager.
- **IBM SPSS Modeler CLEF Guide du développeur.** CLEF permet d'intégrer des programmes tiers tels que des programmes de traitement de données ou des algorithmes de modélisation en tant que noeuds dans SPSS Modeler.
- **Guide d'exploration de base de données IBM SPSS Modeler.** Informations sur la manière de tirer parti de la puissance de votre base de données pour améliorer les performances et étendre la gamme des fonctions analytiques via des algorithmes tiers.
- **Guide des performances et d'administration IBM SPSS Modeler Server.** Informations sur le mode de configuration et d'administration de IBM® SPSS® Modeler Server.
- **Guide de l'utilisateur de IBM SPSS Modeler Administration Console.** Informations concernant l'installation et l'utilisation de l'interface utilisateur de la console permettant de surveiller et de configurer SPSS Modeler Server. La console est implémentée en tant que plug-in à l'application Deployment Manager.
- **Guide IBM SPSS Modeler Solution Publisher.** SPSS Modeler Solution Publisher est un module complémentaire qui permet aux entreprises de publier des flux destinés à être utilisés en dehors de l'environnement SPSS Modeler.
- **Guide CRISP-DM IBM SPSS Modeler** Guide détaillé sur l'utilisation de la méthodologie CRISP-DM pour le Data mining avec SPSS Modeler
- **Guide de l'utilisateur IBM SPSS Modeler Batch.** Guide complet sur l'utilisation de IBM SPSS Modeler en mode par lots, avec des détails sur l'exécution en mode par lots et les arguments de ligne de commande. Ce guide est disponible au format PDF uniquement.

## ***Documentation de SPSS Modeler Premium***

La suite de documentation SPSS Modeler Premium (à l'exception des instructions d'installation) est la suivante.

- **IBM SPSS Modeler Entity Analytics Guide de l'utilisateur.** Informations sur l'utilisation des analyses d'entités avec SPSS Modeler, notamment l'installation et la configuration du référentiel, les noeuds d'analyses d'entités et les tâches administratives.

- **IBM SPSS Modeler Social Network Analysis Guide de l'utilisateur.** Guide sur l'exécution des analyses de réseaux sociaux avec SPSS Modeler, y compris les analyses de groupe et analyses de diffusion.
- **Text Analytics for SPSS Modeler Guide de l'utilisateur.** Informations sur l'utilisation des analyses de texte avec SPSS Modeler, notamment sur les nœuds de Text Mining, l'espace de travail interactif, les modèles et d'autres ressources.
- Guide de l'utilisateur de **Text Analytics for IBM SPSS Modeler Administration Console.** Informations concernant l'installation et l'utilisation de l'interface utilisateur de la console permettant de surveiller et de configurer IBM® SPSS® Modeler Server pour l'utiliser avec Text Analytics for SPSS Modeler. La console est implémentée en tant que plug-in à l'application Deployment Manager.

## ***Exemples d'application***

Tandis que les outils de Data mining de SPSS Modeler peuvent vous aider à résoudre une grande variété de problèmes commerciaux et organisationnels, les exemples d'application fournissent des introductions brèves et ciblées aux méthodes et aux techniques de modélisation. Les ensembles de données utilisés ici sont beaucoup plus petits que les énormes entrepôts de données gérés par certains Data miners, mais les concepts et les méthodes impliqués doivent pouvoir être adaptés à des applications réelles.

Vous pouvez accéder aux exemples en cliquant Exemples d'application dans le menu Aide de SPSS Modeler. Les fichiers de données et les flux d'échantillons sont installés dans le dossier *Demos*, sous le répertoire d'installation du produit. [Pour plus d'informations, reportez-vous à la section Dossier Demos dans \*Guide de l'utilisateur de IBM SPSS Modeler 15\*.](#)

**Exemples de modélisation de bases de données.** Consultez les exemples dans le *IBM SPSS ModelerGuide d'exploration de base de données*.

**Exemples de génération de scripts.** Consultez les exemples dans le *IBM SPSS ModelerGuide de génération de scripts et d'automatisation*.

## ***Dossier Demos***

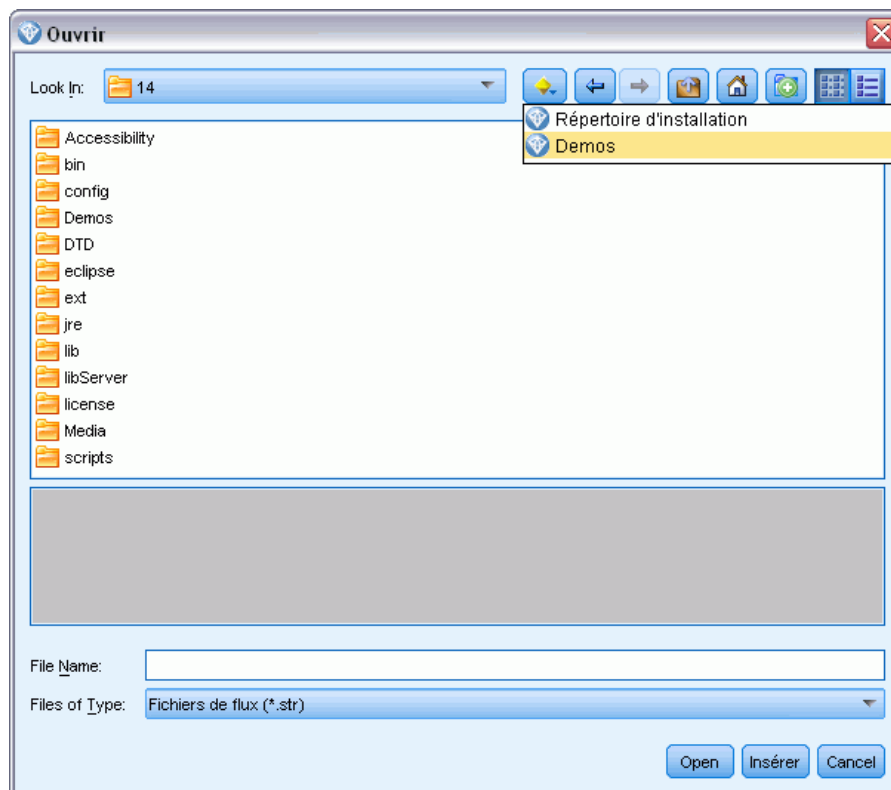
Les fichiers de données et les flux d'échantillons utilisés avec les exemples d'application sont installés dans le dossier *Demos*, sous le répertoire d'installation du produit. Ce dossier est également accessible à partir du groupe de programmes sous IBM SPSS Modeler 15 dans le menu



Démarrer de Windows, ou en cliquant sur *Demos* dans la liste des répertoires récents de la boîte de dialogue Ouverture de fichier.

Figure 1-1

Sélection du dossier *Demos* dans la liste des répertoires récemment consultés



# Exploration des bases de données

## Modélisation de base de données - Présentation

IBM® SPSS® Modeler Server prend en charge l'intégration des outils de Data mining et de modélisation disponibles auprès des fournisseurs de base de données, notamment IBM Netezza, IBM DB2 InfoSphere Warehouse, Oracle Data Miner et Microsoft Analysis Services. Dans la base de données, vous pouvez créer, stocker et déterminer le score des modèles qui proviennent tous de l'application IBM® SPSS® Modeler. Vous pouvez ainsi combiner les capacités analytiques et la facilité d'utilisation de SPSS Modeler avec la puissance et les performances d'une base de données, et ce tout en tirant parti des algorithmes natifs de base de données de ces fournisseurs. Les modèles sont créés à l'intérieur de la base de données. Vous pouvez alors normalement les parcourir et les évaluer via l'interface SPSS Modeler, mais également les déployer en utilisant IBM® SPSS® Modeler Solution Publisher, si besoin est. Les algorithmes pris en charge sont disponibles dans la palette Modélisation de la base de données de SPSS Modeler.

Utiliser SPSS Modeler pour accéder aux algorithmes natifs de base de données présente plusieurs avantages :

- En général, les algorithmes de base de données sont étroitement intégrés au serveur de base de données et peuvent améliorer les performances.
- Les modèles créés et stockés dans une base de données peuvent être facilement déployés vers l'application accédant à cette base de données et partagés avec cette application.

**génération SQL.** La modélisation de base de données est distincte de la génération SQL, appelé également "répercussions SQL". Cette fonctionnalité permet de générer des instructions SQL pour les opérations natives SPSS Modeler qui peuvent être "répercutées" (c'est-à-dire exécutées) dans la base de données afin d'améliorer les performances. Par exemple, les noeuds Fusionner, Agréger et Sélectionner génèrent tous du code SQL pouvant être répercuté de cette manière dans la base de données. Grâce à la combinaison des fonctions de génération SQL et de modélisation de base de données, vous pouvez exécuter les flux de bout en bout dans la base de données. Résultat ? Vous bénéficiez de gains importants en termes de performances, en comparaison des flux exécutés dans SPSS Modeler. [Pour plus d'informations, reportez-vous à la section Optimisation SQL dans le chapitre 6 dans \*Guide d'administration et des performances de IBM SPSS Modeler Server 15\*.](#)

*Remarque :* La modélisation de base de données et l'optimisation SQL requièrent l'activation de la connectivité à SPSS Modeler Server sur l'ordinateur IBM® SPSS® Modeler. Avec ce paramètre activé, vous pouvez accéder aux algorithmes de la base de données, effectuer le push back de SQL directement depuis SPSS Modeler et accéder à SPSS Modeler Server. Pour vérifier l'état actuel de la licence, choisissez ce qui suit dans le menu SPSS Modeler.

Aide > A propos de > Informations supplémentaires

Si la connectivité est activée, vous voyez l'option Activation du serveur dans l'onglet Etat de la licence.

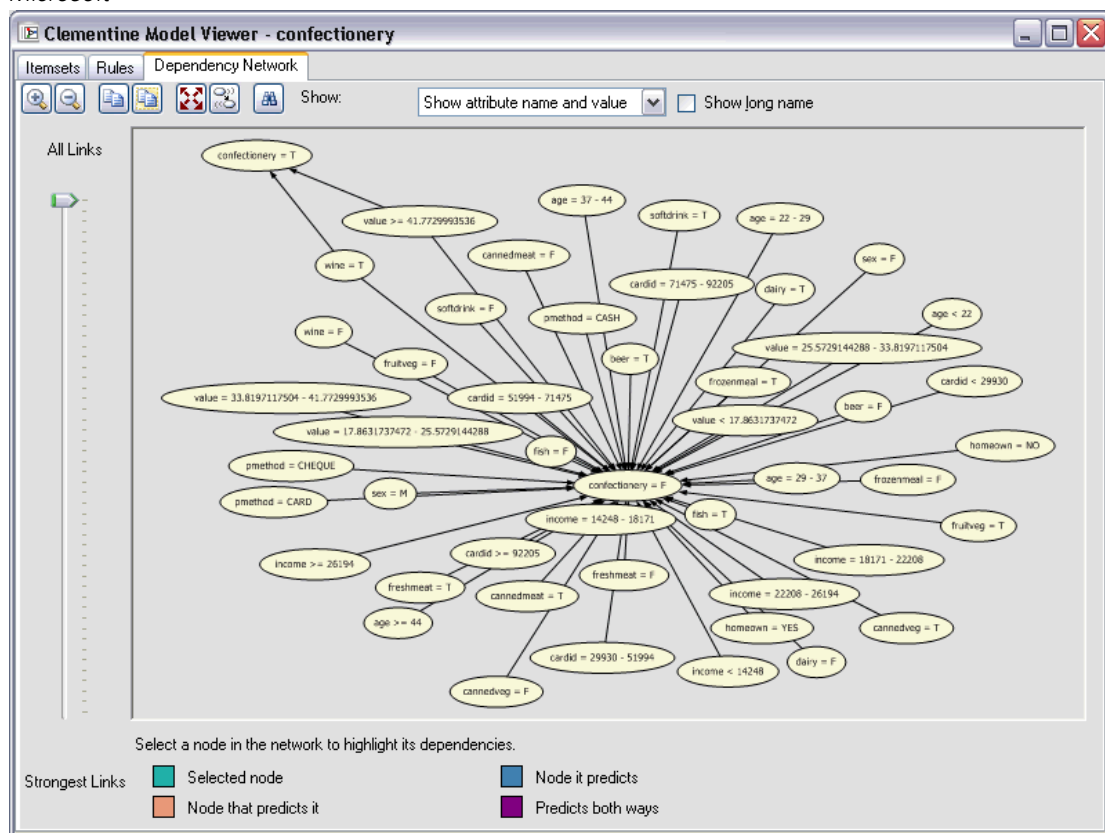
Pour plus d'informations, reportez-vous à la section Connexion au IBM SPSS Modeler Server dans le chapitre 3 dans *Guide de l'utilisateur de IBM SPSS Modeler 15*.

Figure 2-1  
Palette Modélisation de la base de données



Pour plus d'informations sur les algorithmes pris en charge, reportez-vous aux sections relatives à certains fournisseurs ci-après.

Figure 2-2  
Afficheur contenant le graphique des résultats du modèle de règles d'association Analysis Services de Microsoft



## Composants requis

Pour pouvoir modéliser une base de données, vous avez besoin de la configuration suivante :

- Une connexion ODBC à une base de données appropriée, ainsi que les composants analytiques nécessaires installés (Microsoft Analysis Services, Oracle Data Miner ou IBM DB2 InfoSphere Warehouse).

- Dans IBM® SPSS® Modeler, vous devez activer la fonction de modélisation de base de données dans la boîte de dialogue Programmes externes (Outils > Programmes externes).
- Les paramètres Générer SQL et Optimisation SQL doivent être activés dans la boîte de dialogue Options utilisateur de IBM® SPSS® Modeler et sur IBM® SPSS® Modeler Server (en cas d'utilisation). [Pour plus d'informations, reportez-vous à la section Performances/Optimisation dans le chapitre 4 dans \*Guide d'administration et des performances de IBM SPSS Modeler Server 15\*. L'optimisation SQL n'est pas indispensable à la modélisation de base de données, mais elle est vivement recommandée pour des raisons de performances.](#)

*Remarque* : La modélisation de base de données et l'optimisation SQL requièrent l'activation de la connectivité à SPSS Modeler Server sur l'ordinateur SPSS Modeler. Avec ce paramètre activé, vous pouvez accéder aux algorithmes de la base de données, effectuer le push back de SQL directement depuis SPSS Modeler et accéder à SPSS Modeler Server. Pour vérifier l'état actuel de la licence, choisissez ce qui suit dans le menu SPSS Modeler.

Aide > A propos de > Informations supplémentaires

Si la connectivité est activée, vous voyez l'option Activation du serveur dans l'onglet Etat de la licence.

[Pour plus d'informations, reportez-vous à la section Connexion au IBM SPSS Modeler Server dans le chapitre 3 dans \*Guide de l'utilisateur de IBM SPSS Modeler 15\*.](#)

Pour plus d'informations, reportez-vous aux sections relatives à certains fournisseurs ci-après.

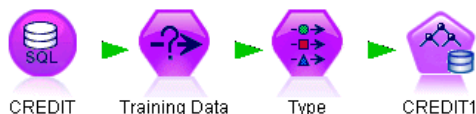
## Construction du modèle

Le processus de création et de scoring de modèles à l'aide d'algorithmes de base de données est similaire aux autres types de Data mining dans IBM® SPSS® Modeler. Dans SPSS Modeler, le processus global permettant de travailler avec les noeuds et de modéliser les « nuggets » est semblable à tout autre flux. Une seule différence est à noter : le traitement réel et la création de modèles sont effectués dans la base de données.

Par exemple, le flux suivant est identique d'un point de vue conceptuel aux autres flux de données dans SPSS Modeler. Il réalise toutefois l'ensemble des opérations dans une base de données, y compris la création de modèles, à l'aide du noeud Arbre décision MS. Lorsque vous exécutez ce flux, SPSS Modeler indique à la base de données de créer et de stocker le modèle obtenu. Les informations correspondantes sont transmises à SPSS Modeler.

Figure 2-3

*Exemple de flux de modélisation de base de données, où les noeuds en violet font référence à l'exécution d'opérations dans une base de données*



## Préparation des données

Que vous utilisiez ou non des algorithmes natifs de base de données, les préparations de données doivent être répercutées dans la base de données chaque fois que cela s'avère possible, afin d'améliorer les performances.

- Si les données d'origine sont stockées dans la base de données, il convient de les y conserver en veillant à ce que toutes les opérations requises effectuées en amont puissent être converties en langage SQL. Vous empêchez ainsi le téléchargement des données vers IBM® SPSS® Modeler—ce qui évite la création d'un goulet d'étranglement qui annulerait toute action bénéfique—et permettez l'exécution de l'intégralité du flux dans la base de données. [Pour plus d'informations, reportez-vous à la section Optimisation SQL dans le chapitre 6 dans \*Guide d'administration et des performances de IBM SPSS Modeler Server 15\*.](#)
- Si les données d'origine ne sont *pas* stockées dans la base de données, vous pouvez toujours utiliser la fonction de modélisation de base de données. Dans ce cas, la préparation des données est effectuée dans SPSS Modeler. En outre, l'ensemble de données préparées est envoyé automatiquement à la base de données pour créer des modèles.

## Scoring de modèles

Les modèles générés depuis IBM® SPSS® Modeler via la fonction d'exploration de base de données sont différents des modèles SPSS Modeler standard. Bien qu'ils apparaissent dans le gestionnaire de modèles en tant que “nuggets,” de modèles générés, ils constituent en réalité des modèles distants gérés sur le serveur de base de données ou de Data mining distant. Les éléments que vous visualisez dans SPSS Modeler ne sont que des références à ces modèles distants. En d'autres termes, le modèle SPSS Modeler affiché est un modèle “vide” contenant différentes informations, comme les noms de l'hôte du serveur de base de données, de la base de données et du modèle. Il s'agit là d'une distinction fondamentale que vous devez assimiler lorsque vous parcourez des modèles créés par le biais d'algorithmes natifs de base de données et que vous déterminez un score pour ces modèles.

Figure 2-4  
Modèle généré “nugget”; pour les arbres décision Microsoft



Après avoir créé un modèle, vous pouvez l'ajouter au flux à des fins de scoring, comme tout autre modèle généré dans SPSS Modeler. Toutes les opérations de scoring sont réalisées au sein de la base de données, même si les opérations en amont ne le sont pas. (Si possible et pour améliorer les performances, les opérations en amont peuvent également être répercutées vers la base de données, mais il ne s'agit pas d'une obligation pour que le scoring ait lieu.) Dans la plupart des cas, vous pouvez également parcourir le modèle généré grâce au navigateur standard mis à disposition par le fournisseur de base de données.

Que ce soit pour la navigation ou le scoring, une connexion active au serveur exécutant Oracle Data Miner, IBM DB2 InfoSphere Warehouse ou Microsoft Analysis Services est nécessaire.

### **Affichage des résultats et paramétrage**

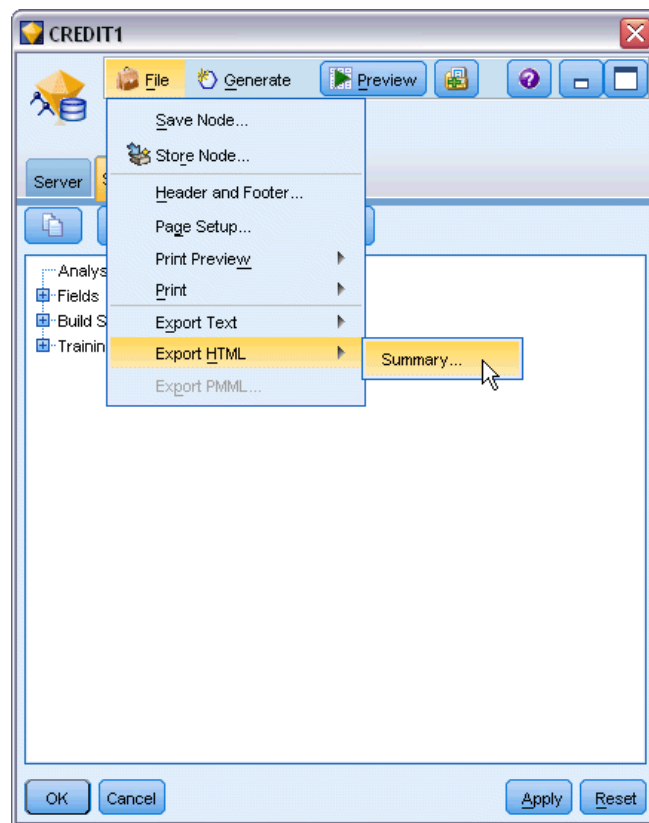
Pour afficher les résultats et indiquer les paramètres de scoring, double-cliquez sur le modèle dans l'espace de travail de flux. Il est également possible de cliquer avec le bouton droit de la souris sur ce modèle, puis de choisir Parcourir ou Edition. Certains paramètres dépendent du type de modèle utilisé.

### **Exportation et enregistrement de modèles de base de données**

Vous pouvez exporter des modèles et résumés de base de données depuis le navigateur de modèle de la même manière que les autres modèles créés dans IBM® SPSS® Modeler, à l'aide des options du menu Fichier.

Figure 2-5

Exportation d'un résumé de modèle d'arbre décision Microsoft au format HTML



- ▶ Dans le menu Fichier du navigateur de modèle, choisissez l'une des options suivantes :
  - Exporter texte exporte le résumé du modèle vers un fichier texte
  - Exporter HTML exporte le résumé du modèle vers un fichier HTML
  - Exporter PMML (pris en charge pour les modèles IBM DB2 IM uniquement) exporte le modèle au format PMML (Predictive Model Markup Language) qui peut être utilisé avec d'autres logiciels compatibles PMML. [Pour plus d'informations, reportez-vous à la section Importation](#)

et exportation de modèles au format PMML dans le chapitre 10 dans *Guide de l'utilisateur de IBM SPSS Modeler 15*.

*Remarque* : Pour enregistrer un modèle généré, vous pouvez également choisir Enregistrer le noeud dans le menu Fichier. Pour plus d'informations, reportez-vous à la section Navigation dans les nuggets de modèle dans le chapitre 3 dans *Noeuds de modélisation de IBM SPSS Modeler 15*.

## Cohérence d'un modèle

IBM® SPSS® Modeler stocke la description de la structure de chaque modèle de base de données généré, ainsi qu'une référence au modèle de même nom stocké dans la base de données. L'onglet Serveur du modèle créé fournit une clé unique générée pour le modèle, qui correspond au modèle réel dans la base de données.

Figure 2-6  
Clé de modèle créée et options de vérification

The screenshot shows the 'Serveur' tab in the IBM SPSS Modeler interface. It contains the following fields and buttons:

- Hôte du serveur d'analyse:** gb1w2k3dbhost1
- Base de données du serveur d'analyse:** dan (with a blue button containing three dots to the right)
- Connexion du serveur SQL:** SQL Server Stndrd (with a blue button containing three dots to the right)
- GUID du modèle:** {2D5DB0DF-5888-43EC-BBC2-1218F057F7AD}
- At the bottom, there are two buttons: 'Vérifier' (with a checkmark icon) and 'Affichage' (with a document icon).

SPSS Modeler utilise cette clé générée de façon aléatoire pour s'assurer que le modèle est toujours cohérent. Cette clé est stockée dans la description du modèle lorsqu'il est créé. Il s'avère judicieux de vérifier la correspondance des clés avant d'exécuter un flux de déploiement.

- Cliquez sur le bouton Vérifier pour vérifier la cohérence du modèle stocké dans la base de données, en comparant sa description à la clé aléatoire stockée par SPSS Modeler. Si le modèle figurant dans la base de données s'avère introuvable ou que la clé ne correspond pas, une erreur est signalée.

## Affichage et exportation de code SQL généré

Vous pouvez prévisualiser le code SQL généré avant de l'exécuter, ce qui peut s'avérer utile pour le débogage. Pour plus d'informations, reportez-vous à la section Prévisualisation du code SQL généré dans le chapitre 6 dans *Guide d'administration et des performances de IBM SPSS Modeler Server 15*.

# ***Modélisation de base de données à l'aide de Microsoft Analysis Services***

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

IBM® SPSS® Modeler prend en charge l'intégration à Microsoft SQL Server Analysis Services. Cette fonction, disponible dans la palette Modélisation de la base de données, est implémentée en tant que noeuds de modélisation dans SPSS Modeler. Si la palette n'est pas visible, vous pouvez l'afficher en activant l'intégration MS Analysis Services (dans l'onglet Microsoft de la boîte de dialogue Programmes externes). [Pour plus d'informations, reportez-vous à la section Activation de l'intégration à Analysis Services sur p. 17.](#)

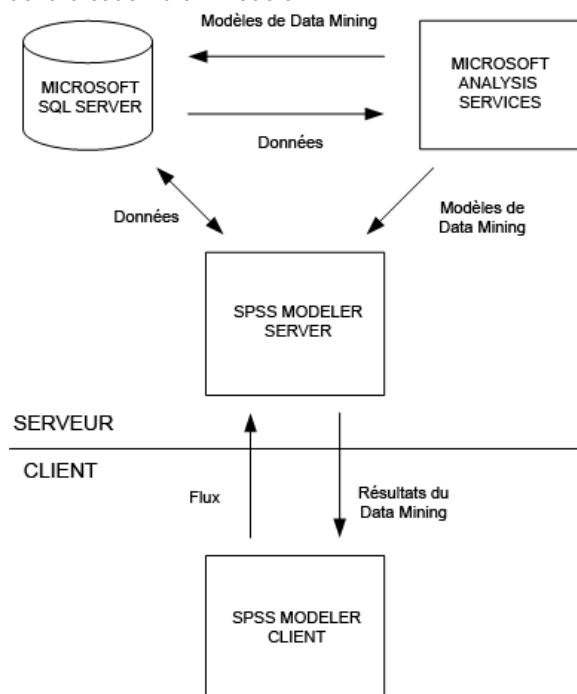
SPSS Modeler prend en charge l'intégration des algorithmes Analysis Services suivants :

- Arbres décision
- Classification
- Règles d'association
- Naive Bayes
- Régression linéaire
- Réseau de neurones
- Régression logistique
- Séries temporelles
- Classification de séquences

Le diagramme suivant illustre le flux de données du client vers le serveur sur lequel IBM® SPSS® Modeler Server gère l'exploration des bases de données. Les modèles sont créés via Analysis Services. Analysis Services stocke le modèle résultant. Une référence à ce modèle est mise à jour dans les flux SPSS Modeler. Ce modèle est ensuite téléchargé depuis Analysis Services vers Microsoft SQL Server ou vers SPSS Modeler à des fins de scoring.



**Figure 3-1**  
Flux de données entre IBM SPSS Modeler, Microsoft SQL Server et Microsoft Analysis Services lors de la création d'un modèle



*Remarque* : Le SPSS Modeler Server n'est pas nécessaire, bien qu'il puisse être utilisé. Le client de IBM® SPSS® Modeler est capable d'effectuer des calculs d'exploration des bases de données lui-même.

### **Conditions requises pour l'intégration à Microsoft Analysis Services**

Les conditions préalables de modélisation de base de données via des algorithmes Analysis Services dans IBM® SPSS® Modeler sont décrites ci-après. N'hésitez pas à contacter l'administrateur de base de données pour vous assurer que ces conditions sont remplies.

- IBM® SPSS® Modeler s'exécutant avec une installation IBM® SPSS® Modeler Server (en mode réparti) sous Windows. Les plates-formes UNIX ne sont pas prises en charge dans cette intégration à Analysis Services.

*Important* : Les utilisateurs de SPSS Modeler doivent configurer une connexion ODBC à l'aide du pilote SQL Native Client disponible auprès de Microsoft à l'URL répertorié ci-dessous dans *Éléments de configuration supplémentaires de SPSS Modeler Server*. Le pilote fourni avec IBM® SPSS® Data Access Pack (généralement conseillé pour toute autre utilisation de SPSS Modeler) n'est pas recommandé dans le cas présent. Vous devez configurer le pilote de manière à ce qu'il utilise SQL Server avec l'authentification Windows intégrée activée ; en effet, SPSS Modeler ne prend pas en charge l'authentification SQL Server. Si vous avez des questions sur la création ou la définition d'autorisations pour les sources de données ODBC, contactez l'administrateur de votre base de données.

- Vous devez installer SQL Server 2005 ou 2008 mais pas obligatoirement sur le même hôte que SPSS Modeler. Les utilisateurs de SPSS Modeler doivent posséder des droits d'accès suffisants pour pouvoir lire et écrire des données, ainsi que supprimer et créer des tables et des vues.

*Remarque* : Il est conseillé d'utiliser SQL Server Enterprise Edition. L'édition Enterprise Edition fournit une flexibilité supplémentaire grâce à des paramètres avancés pour régler les résultats des algorithmes. L'édition Standard Edition fournit les mêmes paramètres mais ne permet pas aux utilisateurs de modifier certains des paramètres avancés.

- Vous devez installer les Composants Analysis Services de Microsoft SQL Server sur le même hôte que SQL Server.

### **Autres éléments IBM SPSS Modeler Server requis**

Si vous souhaitez utiliser les algorithmes Analysis Services avec SPSS Modeler Server, l'ordinateur hôte SPSS Modeler Server doit disposer des composants suivants.

*Remarque* : Si SQL Server est déjà installé sur le même hôte que SPSS Modeler Server, les composants en question le sont également.

- Package redistribuable de Microsoft .NET Framework version 2.0 (x86)
- Microsoft Core XML Services (MSXML) 6.0
- Fournisseur OLE DB pour Microsoft SQL Server 2008 Analysis Services 10.0 (assurez-vous de sélectionner la variante appropriée à votre système d'exploitation)
- Microsoft SQL Server 2008 Native Client (assurez-vous de sélectionner la variante appropriée à votre système d'exploitation)

Pour télécharger ces composants, rendez-vous sur [www.microsoft.com/downloads](http://www.microsoft.com/downloads), recherchez .NET Framework ou (pour tous les autres composants) SQL Server Feature Pack, et sélectionnez le dernier pack pour votre version de SQL Server.

Ces composants peuvent nécessiter l'installation préalable d'autres logiciels, également disponibles sur le site Web des téléchargements de Microsoft.

### **Autres éléments IBM SPSS Modeler requis**

Pour pouvoir utiliser les algorithmes Analysis Services avec SPSS Modeler, les mêmes composants que ceux indiqués ci-dessus doivent être installés, en plus des composants suivants sur le client :

- Microsoft SQL Server 2008 Datamining Viewer Controls (assurez-vous de sélectionner la variante appropriée à votre système d'exploitation). Vous aurez également besoin de :
- Microsoft ADOMD.NET

Pour télécharger ces composants, rendez-vous sur [www.microsoft.com/downloads](http://www.microsoft.com/downloads), recherchez SQL Server Feature Pack, et sélectionnez le dernier pack pour votre version de SQL Server.

*Remarque* : La modélisation de base de données et l'optimisation SQL requièrent l'activation de la connectivité à SPSS Modeler Server sur l'ordinateur SPSS Modeler. Avec ce paramètre activé, vous pouvez accéder aux algorithmes de la base de données, effectuer le push back de SQL

directement depuis SPSS Modeler et accéder à SPSS Modeler Server. Pour vérifier l'état actuel de la licence, choisissez ce qui suit dans le menu SPSS Modeler.

Aide > A propos de > Informations supplémentaires

Si la connectivité est activée, vous voyez l'option Activation du serveur dans l'onglet Etat de la licence.

Pour plus d'informations, reportez-vous à la section Connexion au IBM SPSS Modeler Server dans le chapitre 3 dans *Guide de l'utilisateur de IBM SPSS Modeler 15*.

## **Activation de l'intégration à Analysis Services**

Pour activer l'intégration de IBM® SPSS® Modeler à Analysis Services, vous devez configurer SQL Server et Analysis Services, créer une source de données ODBC, activer cette intégration dans la boîte de dialogue Programmes externes de SPSS Modeler, puis activer la génération et l'optimisation SQL .

*Remarque* : Microsoft SQL Server et Microsoft Analysis Services doivent être disponibles. [Pour plus d'informations, reportez-vous à la section Conditions requises pour l'intégration à Microsoft Analysis Services sur p. 15.](#)

### **Configuration de SQL Server**

Configurez SQL Server pour que le scoring puisse avoir lieu dans la base de données.

- ▶ Créez la clé de registre suivante sur l'ordinateur hôte SQL Server :

HKEY\_LOCAL\_MACHINE\SOFTWARE\Microsoft\MSSQLServer\Providers\MSOLAP

- ▶ Ajoutez à cette clé la valeur DWORD suivante :

AllowInProcess 1

- ▶ Redémarrez SQL Server une fois cette modification apportée.

### **Configuration d'Analysis Services**

Pour que SPSS Modeler puisse communiquer avec Analysis Services, vous devez d'abord configurer manuellement deux paramètres de la boîte de dialogue des propriétés du serveur d'analyse :

- ▶ Connectez-vous au serveur d'analyse via MS SQL Server Management Studio.
- ▶ Accédez à la boîte de dialogue des propriétés en cliquant avec le bouton droit de la souris sur le nom du serveur et en sélectionnant Propriétés.
- ▶ Cochez la case Afficher (toutes) les propriétés avancées.

- ▶ Modifiez les propriétés suivantes :
  - Affectez la valeur `True` (vrai) à `DataMining\AllowAdHocOpenRowsetQueries` (la valeur par défaut est `False`(faux)).
  - Affectez la valeur `[all]` à `DataMining\AllowProvidersInOpenRowset` (aucune valeur par défaut n'est définie).

### **Création d'un DSN ODBC pour SQL Server**

Pour lire ou écrire sur une base de données, vous devez installer et configurer une source de données ODBC pour la base de données appropriée, avec, le cas échéant, des autorisations en lecture et en écriture. Le pilote ODBC pour Microsoft SQL Native Client est requis et est installé automatiquement avec SQL Server. *Le pilote fourni avec IBM® SPSS® Data Access Pack (généralement conseillé pour toute autre utilisation de SPSS Modeler) n'est pas recommandé dans le cas présent.* Si SPSS Modeler et SQL Server se trouvent sur différents hôtes, vous pouvez télécharger le pilote ODBC pour Microsoft SQL Native Client. [Pour plus d'informations, reportez-vous à la section Conditions requises pour l'intégration à Microsoft Analysis Services sur p. 15.](#)

Si vous avez des questions sur la création ou la définition d'autorisations pour les sources de données ODBC, contactez l'administrateur de votre base de données.

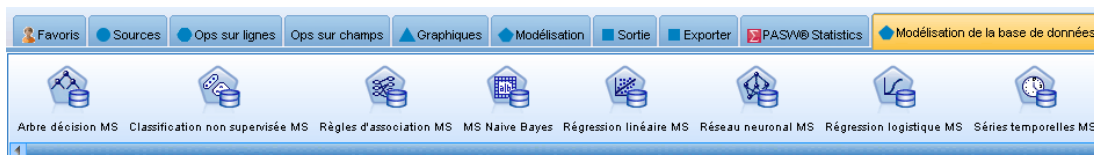
- ▶ A l'aide du pilote ODBC pour Microsoft SQL Native Client, créez un DSN (nom de source de données) ODBC qui pointe vers la base de données SQL Server utilisée dans le processus de Data mining. Conservez les valeurs par défaut des autres paramètres du pilote.
- ▶ Pour ce DSN, assurez-vous que l'option Avec authentification Windows intégrée est sélectionnée.
  - Si IBM® SPSS® Modeler et IBM® SPSS® Modeler Server sont exécutés sur des hôtes différents, créez le même DSN ODBC sur chaque hôte. Veillez à utiliser le même nom DSN sur chaque hôte.

### **Activation de l'intégration d'Analysis Services dans IBM SPSS Modeler**

Pour configurer SPSS Modeler de manière à utiliser Analysis Services, indiquez tout d'abord les spécifications du serveur dans la boîte de dialogue Programmes externes.

- ▶ Dans les menus SPSS Modeler, sélectionnez :  
Outils > Options > Programmes externes
- ▶ Cliquez sur l'onglet Microsoft.
  - **Activer l'intégration de Microsoft Analysis Services.** Active la palette Modélisation de la base de données (si elle n'est pas déjà visible) en bas de la fenêtre SPSS Modeler et ajoute les noeuds des algorithmes Analysis Services.

Figure 3-2  
Onglet Modélisation de la base de données



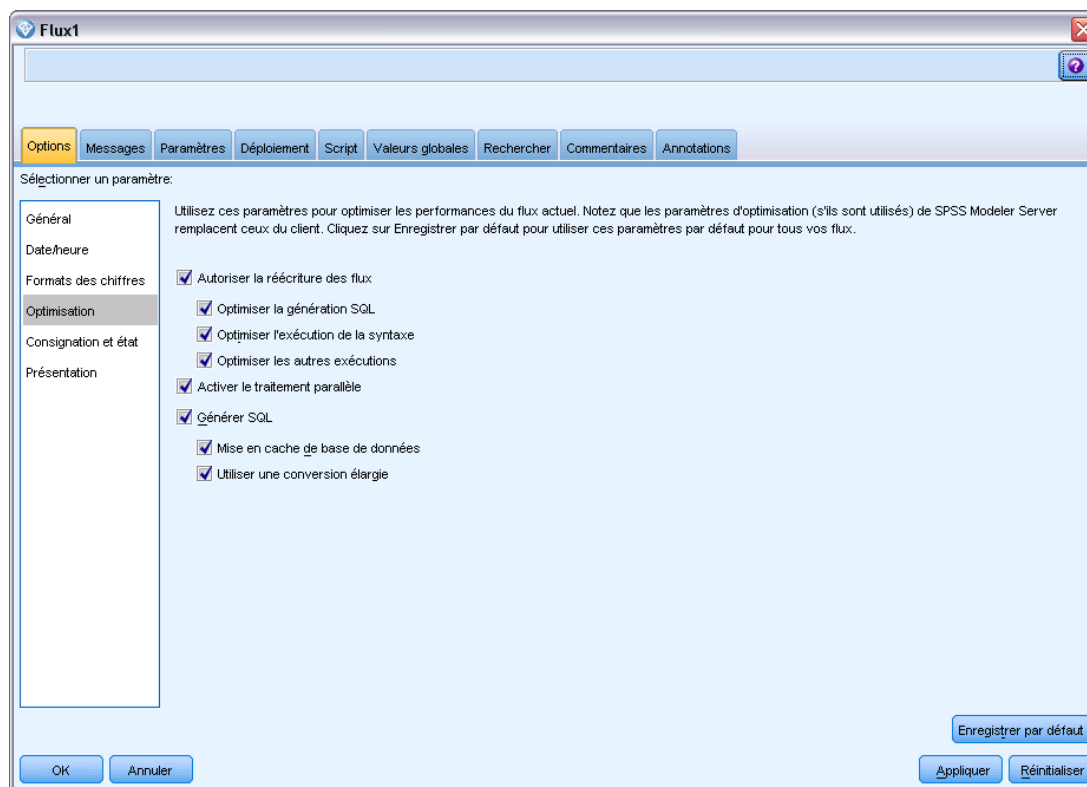
- **Hôte du serveur d'analyse.** Spécifiez le nom de l'ordinateur sur lequel Analysis Services est exécuté.
- **Base de données du serveur d'analyse.** Sélectionnez la base de données voulue en cliquant sur le bouton représentant des points de suspension (...) pour ouvrir une sous-boîte de dialogue. Dans cette dernière, vous pouvez choisir parmi les bases de données disponibles. La liste contient les bases de données disponibles pour le serveur d'analyse indiqué. Microsoft Analysis Services stockant les modèles de Data mining dans des bases de données nommées, vous devez sélectionner la base de données appropriée, qui contient les modèles Microsoft générés par SPSS Modeler.
- **Connexion du serveur SQL.** Fournissez les informations de DSN utilisées par la base de données SQL Server pour stocker les données qui sont transmises au serveur d'analyse. Sélectionnez la source de données ODBC qui permet de fournir les données nécessaires à la création de modèles de Data mining Analysis Services. Si vous générez des modèles Analysis Services à partir des données incluses dans les fichiers plats ou les sources de données ODBC, ces données sont envoyées automatiquement à une table temporaire créée dans la base de données SQL Server vers laquelle pointe la source de données ODBC.
- **Avertir au moment de remplacer un modèle de Data mining.** Sélectionnez cette option pour vous assurer que les modèles stockés dans la base de données ne sont pas ignorés par SPSS Modeler sans que vous en soyez informé.

*Remarque :* Les paramètres définis dans la boîte de dialogue Programmes externes peuvent être ignorés dans les divers noeuds Analysis Services.

#### **Activation de la génération SQL et de l'optimisation**

- ▶ Dans les menus SPSS Modeler, sélectionnez :  
Outils > Propriétés du flux > Options

Figure 3-3  
Paramètres d'optimisation



- ▶ Cliquez sur l'option Optimisation dans le volet de navigation.
- ▶ Confirmez que l'option Générer SQL est bien activée. Ce paramètre doit être utilisé pour que la modélisation de base de données puisse fonctionner.
- ▶ Sélectionnez Optimiser la génération SQL et Optimiser les autres exécutions (cette opération n'est pas forcément nécessaire, mais elle est vivement recommandée pour optimiser les performances).

Pour plus d'informations, reportez-vous à la section Définition des options d'optimisation pour les flux dans le chapitre 5 dans *Guide de l'utilisateur de IBM SPSS Modeler 15*.

## Création de modèles à l'aide d'Analysis Services

La création de modèles avec Analysis Services requiert la présence de l'ensemble de données d'apprentissage dans une table ou une vue de la base de données SQL Server. Si les données ne se trouvent pas dans SQL Server ou doivent être traitées dans IBM® SPSS® Modeler, dans le cadre de la préparation des données ne pouvant pas être exécutées dans SQL Server, les données sont automatiquement envoyées vers une table temporaire dans SQL Server avant la création du modèle.

## Gestion des modèles Analysis Services

Générer un modèle Analysis Services via IBM® SPSS® Modeler crée un modèle dans SPSS Modeler, et génère ou remplace un modèle dans la base de données SQL Server. Le modèle SPSS Modeler fait référence au contenu d'un modèle de base de données stocké sur un serveur de base de données. SPSS Modeler peut vérifier la cohérence en stockant une chaîne de clé de modèle généré identique dans les modèles SPSS Modeler et SQL Server.



Le noeud de modélisation **Arbre décision MS** est utilisé dans la modélisation prédictive des attributs catégoriels et continus. Pour les attributs catégoriels, le noeud réalise des prévisions sur la base des relations entre les différentes colonnes d'entrée d'un ensemble de données. Supposons, par exemple, que vous souhaitiez déterminer les clients susceptibles d'acheter un vélo. Si neuf jeunes clients sur dix achètent un vélo, contre seulement deux clients plus âgés sur dix, le noeud en déduit que l'âge est un facteur important dans l'acquisition d'un vélo. L'arbre décision réalise ainsi des prévisions en fonction de la tendance d'un résultat particulier. Pour les attributs continus, l'algorithme utilise la régression linéaire afin de déterminer l'emplacement où l'arbre décision se divise. Si plusieurs colonnes sont définies comme étant prévisibles ou si les données d'entrée contiennent une table imbriquée définie comme étant prévisible, le noeud crée un arbre décision distinct pour chaque colonne prévisible.



Le noeud de modélisation **Classification non supervisée MS** utilise des techniques itératives pour regrouper les observations d'un ensemble de données dans des classes présentant des caractéristiques semblables. Ces regroupements s'avèrent utiles pour l'exploration des données, l'identification des anomalies au sein des données et la création de prévisions. Les modèles de classification non supervisée identifient les relations d'un ensemble de données, relations que vous ne pourriez peut-être pas calculer de manière logique par simple observation. Par exemple, vous pouvez déduire de manière logique que les personnes qui vont travailler à vélo n'habitent généralement pas trop loin de leur lieu de travail. L'algorithme peut toutefois identifier d'autres caractéristiques moins évidentes sur les personnes allant travailler à vélo. Le noeud Classification non supervisée diffère des autres noeuds de Data mining en ce sens qu'aucun champ cible n'est indiqué pour lui. Il forme le modèle strictement à partir des relations qui existent au sein des données et à partir des classes identifiées par le noeud.



Le noeud de modélisation **Règles d'association Microsoft** est utile pour les moteurs de recommandations. Un moteur de recommandations conseille des produits aux clients en fonction des articles que ceux-ci ont déjà achetés ou pour lesquels ils ont manifesté un intérêt. Les modèles d'association sont créés à partir d'ensembles de données contenant des identificateurs pour chaque observation ainsi que pour les éléments que comportent ces observations. Un groupe d'éléments contenu dans une observation est appelé **jeu d'éléments**. Un modèle d'association est composé d'une série de jeux d'éléments, ainsi que des règles qui décrivent le mode selon lequel les éléments sont regroupés aux seins des observations. Les règles identifiées par l'algorithme peuvent être utilisées pour prévoir les futurs achats possibles d'un client, sur la base des éléments déjà présents dans son panier.



Le noeud de modélisation **MS Naive Bayes** calcule la probabilité conditionnelle entre les champs cible et les champs prédicteurs, et suppose que les colonnes sont indépendantes. Ce modèle est appelé naïve, car il considère toutes les variables de prévision proposées comme étant indépendantes les unes des autres. Cette méthode est moins poussée en termes de calcul que les autres algorithmes Analysis Services et peut donc être utile pour rechercher rapidement des relations lors des étapes préliminaires de modélisation. Vous pouvez utiliser ce noeud pour réaliser une exploration initiale des données, puis exploiter les résultats afin de créer des modèles supplémentaires avec d'autres noeuds qui effectuent des calculs plus longs mais fournissent des résultats plus précis.



Le noeud de modélisation **Régression linéaire MS** est une variante du noeud Arbres de décision, où le paramètre `MINIMUM_LEAF_CASES` est défini comme étant supérieur ou égal au nombre total de cas, dans l'ensemble de données qu'utilise le noeud pour former le modèle de Data mining. Ce paramètre ainsi défini, le noeud ne crée jamais de division et exécute alors une régression linéaire.



Le noeud de modélisation **Réseau de neurones MS** est semblable au noeud Arbre de décision MS car il calcule les probabilités pour chaque état possible de l'attribut d'entrée lorsque chaque état de l'attribut prévisible est donné. Vous pouvez ultérieurement utiliser ces probabilités pour prédire un résultat de l'attribut prédit en fonction des attributs d'entrée.



Le noeud de modélisation **Régression logistique MS** est une variante du noeud R. neurones MS, où la valeur du paramètre `HIDDEN_NODE_RATIO` est 0. Ce paramètre crée un modèle de réseau de neurones exempt de couche cachée et qui équivaut par conséquent à la régression logistique.



Le noeud de modélisation des **séries temporelles MS** offre des algorithmes de régression qui sont optimisés pour les prévisions des valeurs continues, comme les ventes de produits, sur une certaine période de temps. Alors que les algorithmes Microsoft, comme les arbres de décision, nécessitent des colonnes supplémentaires pour y saisir de nouvelles informations permettant de prédire une tendance, un modèle de séries temporelles n'en a pas besoin. Un modèle de séries temporelles peut prédire des tendances uniquement en fonction de l'ensemble de données d'origine qui est utilisé pour créer le modèle. Vous pouvez également ajouter de nouvelles données au modèle lorsque vous effectuez une prédiction et incorporer automatiquement ces nouvelles données dans l'analyse des tendances. [Pour plus d'informations, reportez-vous à la section Noeud Séries temporelles MS sur p. 32.](#)



Le noeud de modélisation **Classification de séquences MS** identifie les séquences ordonnées en données, et combine les résultats de cette analyse avec des techniques de classification pour générer des classes basées sur les séquences et les autres attributs. [Pour plus d'informations, reportez-vous à la section Noeud Classification de séquences MS sur p. 36.](#)

Vous pouvez accéder à chaque noeud depuis la palette Modélisation de la base de données située au bas de la fenêtre SPSS Modeler.

### **Paramètres communs à tous les noeuds d'algorithme**

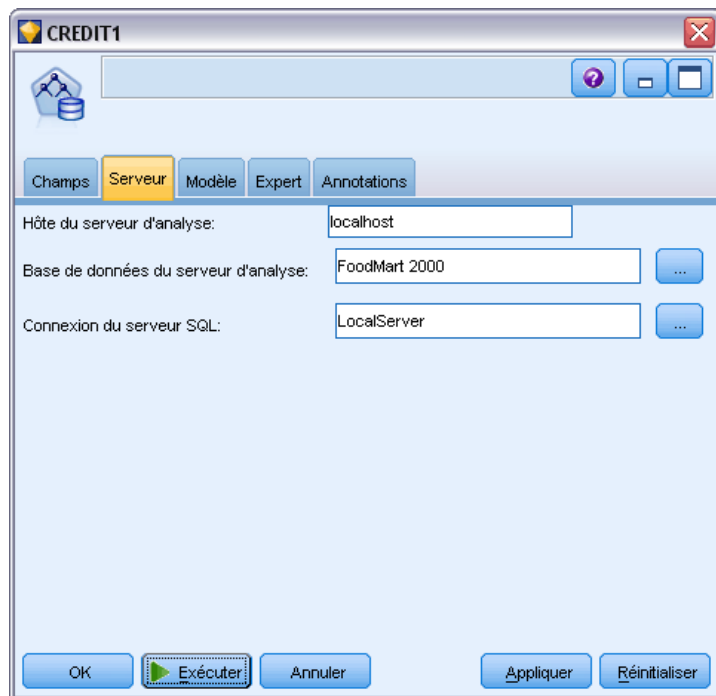
Les valeurs suivantes sont communes à tous les algorithmes Analysis Services.



### Options du serveur

Dans l'onglet Serveur, vous pouvez configurer l'hôte et la base de données du serveur d'analyse, ainsi que la source de données SQL Server. Les options spécifiées dans cet onglet remplacent celles figurant dans l'onglet Microsoft de la boîte de dialogue Programmes externes. [Pour plus d'informations, reportez-vous à la section Activation de l'intégration à Analysis Services sur p. 17.](#)

Figure 3-4  
Options du serveur

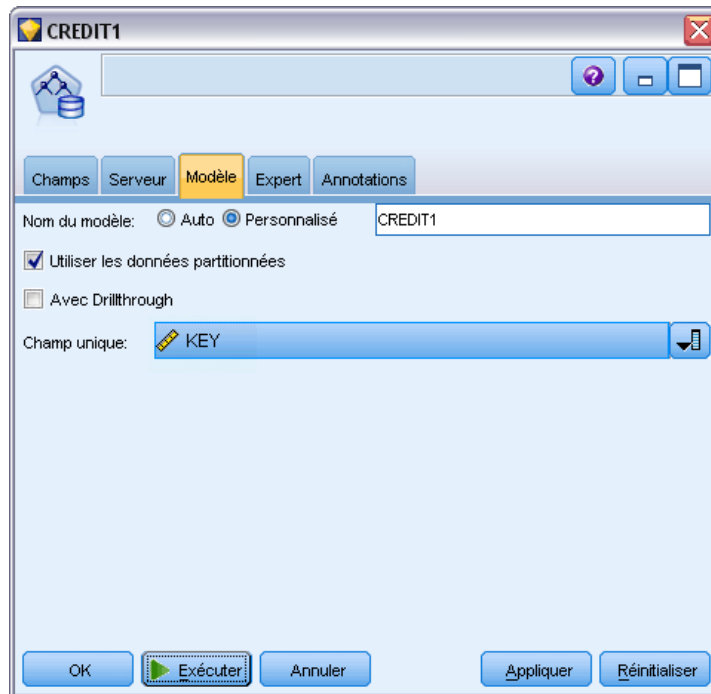


*Remarque* : Un onglet du même type est également disponible lors du scoring des modèles Analysis Services. [Pour plus d'informations, reportez-vous à la section Onglet Serveur du nugget de modèle Analysis Services sur p. 39.](#)

### Options de modèle

Pour pouvoir créer le modèle le plus simple, vous devez indiquer les options correspondantes dans l'onglet Modèle avant de continuer. La méthode de scoring et les autres options avancées sont disponibles dans l'onglet Expert.

Figure 3-5  
Options de modèle



Les options suivantes de modélisation de base sont disponibles :

**Nom du modèle.** Indique le nom attribué au modèle créé lors de l'exécution du noeud.

- **Auto :** Génère automatiquement le nom du modèle en fonction du nom du champ cible ou du champ d'ID, ou en fonction du nom du type du modèle si aucune cible n'est précisée (comme c'est le cas des modèles de classe).
- **Personnalisée.** Permet de donner un nom personnalisé au modèle créé.

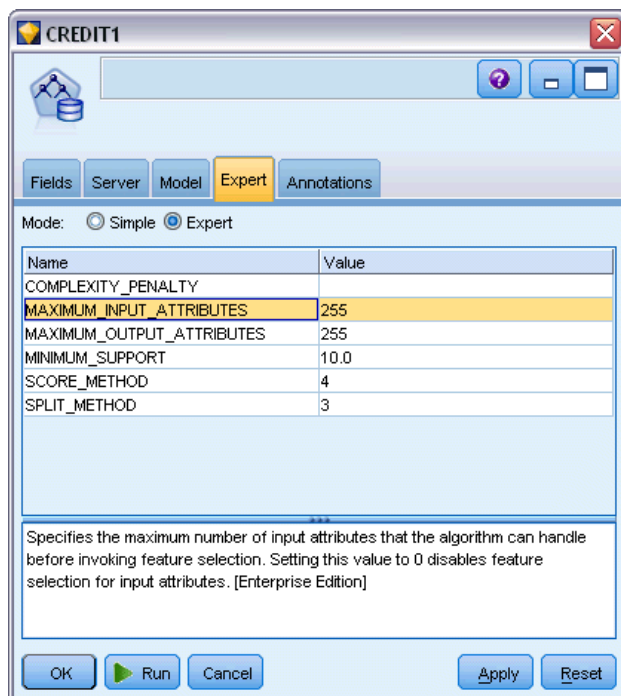
**Utiliser les données partitionnées.** Divise les données en sous-ensembles ou en échantillons distincts destinés à la formation, au test et à la validation, en fonction du champ de partitionnement actuel. L'utilisation d'un échantillon pour la création du modèle et d'un échantillon distinct pour le tester est un indicateur de la manière dont le modèle peut se généraliser à des ensembles de données plus importants, similaires aux données actuelles. Si aucun champ de partitionnement n'est indiqué dans le flux, cette option n'est pas prise en compte. [Pour plus d'informations, reportez-vous à la section Noeud Partitionner dans le chapitre 4 dans Noeuds source, exécution et de sortie de IBM SPSS Modeler 15.](#)

**Avec extraction.** Si elle apparaît, cette option vous permet d'interroger le modèle pour obtenir des détails sur les observations incluses dans le modèle.

**Champ unique.** Dans la liste déroulante, sélectionnez un champ identifiant chaque observation de façon unique. En général, il s'agit d'un champ d'ID, tel que ID client.

## Arbre de décision MS - Options expert

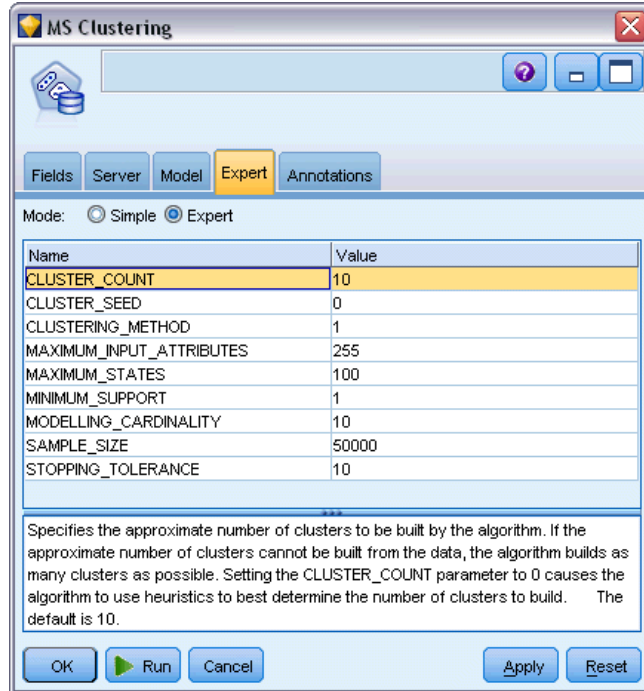
Figure 3-6  
Arbre de décision MS - Options expert



Les options disponibles dans l'onglet Expert peuvent varier en fonction de la structure du flux sélectionné. Consultez l'aide par champ de l'interface utilisateur pour obtenir des détails complets sur les options expert disponibles pour le noeud de modèle Analysis Services sélectionné.

## Classification MS - Options expert

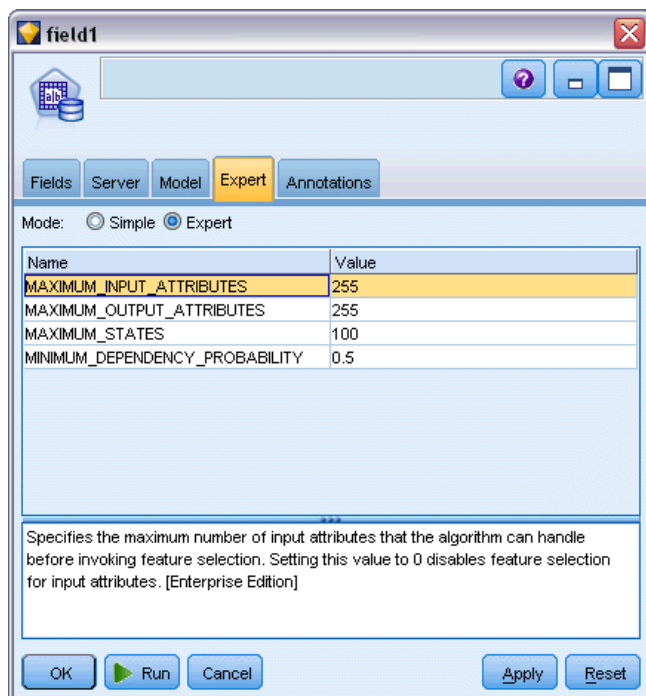
Figure 3-7  
Classification MS - Options expert



Les options disponibles dans l'onglet Expert peuvent varier en fonction de la structure du flux sélectionné. Consultez l'aide par champ de l'interface utilisateur pour obtenir des détails complets sur les options expert disponibles pour le noeud de modèle Analysis Services sélectionné.

## Options expert MS Naive Bayes

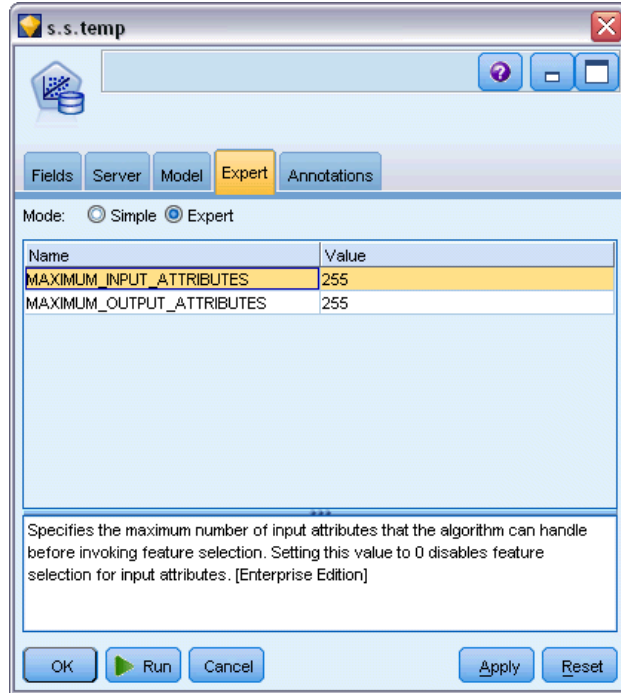
Figure 3-8  
Options expert MS Naive Bayes



Les options disponibles dans l'onglet Expert peuvent varier en fonction de la structure du flux sélectionné. Consultez l'aide par champ de l'interface utilisateur pour obtenir des détails complets sur les options expert disponibles pour le noeud de modèle Analysis Services sélectionné.

## Régression linéaire MS - Options expert

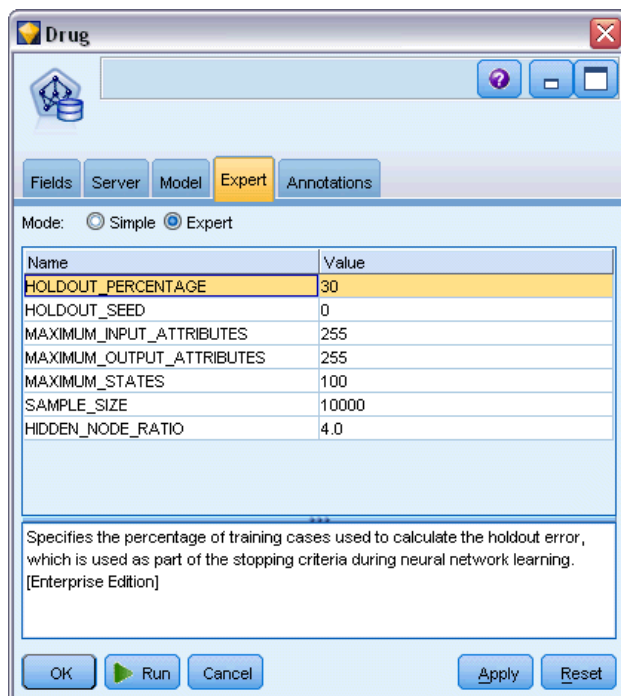
Figure 3-9  
Régression linéaire MS - Options expert



Les options disponibles dans l'onglet Expert peuvent varier en fonction de la structure du flux sélectionné. Consultez l'aide par champ de l'interface utilisateur pour obtenir des détails complets sur les options expert disponibles pour le noeud de modèle Analysis Services sélectionné.

## Réseau de neurones MS - Options expert

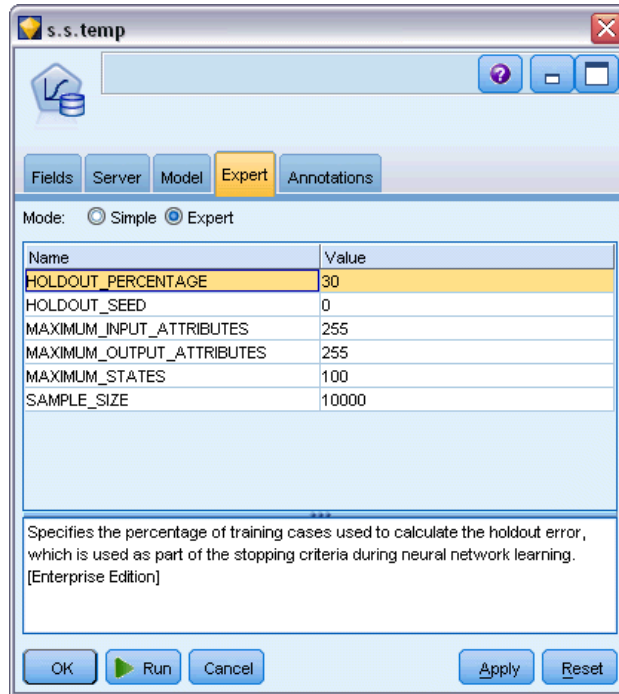
Figure 3-10  
Réseau de neurones MS - Options expert



Les options disponibles dans l'onglet Expert peuvent varier en fonction de la structure du flux sélectionné. Consultez l'aide par champ de l'interface utilisateur pour obtenir des détails complets sur les options expert disponibles pour le noeud de modèle Analysis Services sélectionné.

## Régression logistique MS - Options expert

Figure 3-11  
Régression logistique MS - Options expert



Les options disponibles dans l'onglet Expert peuvent varier en fonction de la structure du flux sélectionné. Consultez l'aide par champ de l'interface utilisateur pour obtenir des détails complets sur les options expert disponibles pour le noeud de modèle Analysis Services sélectionné.

## Noeud Règles d'association MS

Le noeud de modélisation Règles d'association MS est utile pour les moteurs de recommandations. Un moteur de recommandations conseille des produits aux clients en fonction des articles que ceux-ci ont déjà achetés ou pour lesquels ils ont manifesté un intérêt. Les modèles d'association sont créés à partir d'ensembles de données contenant des identificateurs pour chaque observation ainsi que pour les éléments que comportent ces observations. Un groupe d'éléments contenu dans une observation est appelé **jeu d'éléments**.

Un modèle d'association est composé d'une série de jeux d'éléments, ainsi que des règles qui décrivent le mode selon lequel les éléments sont regroupés aux seins des observations. Les règles identifiées par l'algorithme peuvent être utilisées pour prévoir les futurs achats possibles d'un client, sur la base des éléments déjà présents dans son panier.

Pour les données au format tabulaire, l'algorithme crée des scores qui représentent les probabilités (\$MP-champ) pour chaque recommandation générée (\$M-champ). Pour les données au format transactionnel, les scores sont créés pour la prise en charge (\$MS-champ), les probabilités (\$MP-champ) et les probabilités ajustées (\$MAP-champ) pour chaque recommandation générée



(*\$M-champ*). Pour plus d'informations, reportez-vous à la section Données tabulaires et données transactionnelles dans le chapitre 12 dans *Noeuds de modélisation de IBM SPSS Modeler 15*.

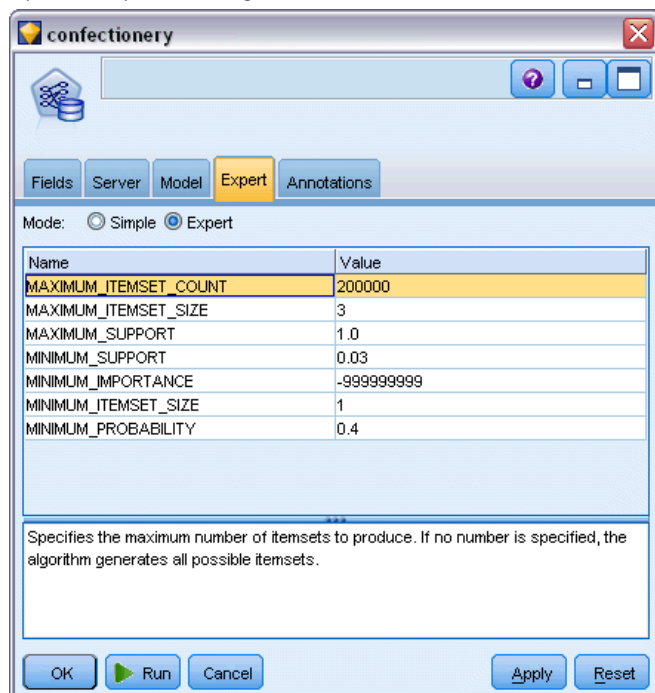
### Conditions requises

La configuration requise pour un modèle d'association transactionnel est la suivante :

- **Champ unique.** Un modèle de règles d'association requiert une clé unique qui identifie les enregistrements.
- **Champ ID.** Lors de la création d'un modèle de règles d'association MS avec des données au format transactionnel, un champ d'ID identifiant chaque transaction est nécessaire. Les champs d'ID peuvent être définis comme le champ unique.
- **Au moins un champ d'entrée.** L'algorithme de règles d'association requiert au moins un champ d'entrée.
- **Champ cible.** Lors de la création d'un modèle d'association MS avec des données transactionnelles, le champ cible doit être le champ transactionnel, par exemple les produits qu'un utilisateur a achetés.

### Options expert des règles d'association MS

Figure 3-12  
Options expert des règles d'association MS



Les options disponibles dans l'onglet Expert peuvent varier en fonction de la structure du flux sélectionné. Consultez l'aide par champ de l'interface utilisateur pour obtenir des détails complets sur les options expert disponibles pour le noeud de modèle Analysis Services sélectionné.

## **Noeud Séries temporelles MS**

Le noeud de modélisation des séries temporelles MS prend en charge deux types de prévisions :

- futures
- historiques

Les **prévisions futures** évaluent les valeurs du champ cible pour un certain nombre de périodes au-delà de la fin de vos données historiques, et celles-ci sont toujours effectuées. Les **prévisions historiques** sont des valeurs de champ cible estimées pour un certain nombre de périodes dont vous avez les vraies valeurs dans vos données historiques. Vous pouvez utiliser les prévisions historiques pour évaluer la qualité du modèle, en comparant les valeurs historiques réelles avec les valeurs prédites. La valeur du point de départ des prévisions détermine si les prédictions historiques seront effectuées.

Contrairement au noeud Séries temporelles de IBM® SPSS® Modeler, il n'est pas nécessaire que le noeud Séries temporelles de MS soit précédé d'un noeud Intervalles de temps. Une autre différence est que par défaut, les scores sont produits uniquement pour les lignes prédites, et non pour toutes les lignes historiques des données de séries temporelles.

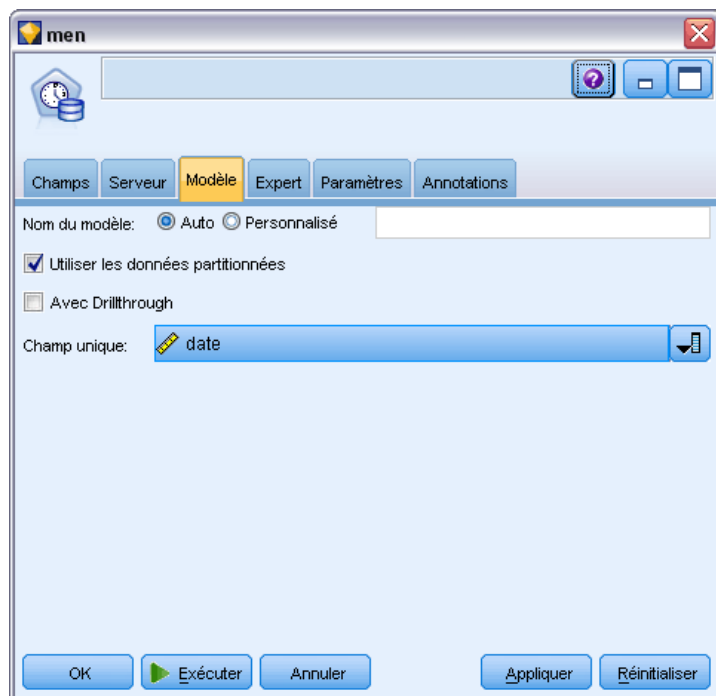
### **Conditions requises**

La configuration requise pour un modèle de séries temporelles MS est la suivante :

- **Champ-clé temporel unique** Chaque modèle doit contenir un champ de date ou un champ numérique utilisé comme série d'observations et qui définit les tranches de temps que le modèle utilisera. Le type de données pour le champ-clé temporel peut être un type de données d'heure ou de date ou un type de données numériques. Cependant, ce champ doit contenir des valeurs continues et ces valeurs doivent être uniques pour chaque série.
- **Champ cible unique.** Vous ne pouvez spécifier qu'un seul champ cible dans chaque modèle. Le type de données du champ cible doit contenir des valeurs continues. Par exemple, vous pouvez prédire comment les attributs numériques, tels que le revenu, les ventes ou la température, évoluent avec le temps. Mais, vous ne pouvez pas utiliser un champ contenant des valeurs catégorielles, comme le statut de l'achat ou le niveau d'éducation, comme champ cible.
- **Au moins un champ d'entrée.** L'algorithme de séries temporelles MS requiert au moins un champ d'entrée. Le type de données du champ d'entrée doit contenir des valeurs continues. Les champs d'entrée non continus sont ignorés pendant la création du modèle.
- **L'ensemble de données doit être trié.** L'ensemble de données d'entrée doit être trié (dans le champ-clé temporel) sinon la création du modèle sera interrompue pour cause d'erreur.

### Options du modèle de séries temporelles MS

Figure 3-13  
Options du modèle de séries temporelles MS



**Nom du modèle.** Indique le nom attribué au modèle créé lors de l'exécution du noeud.

- **Auto :** Génère automatiquement le nom du modèle en fonction du nom du champ cible ou du champ d'ID, ou en fonction du nom du type du modèle si aucune cible n'est précisée (comme c'est le cas des modèles de classe).
- **Personnalisée.** Permet de donner un nom personnalisé au modèle créé.

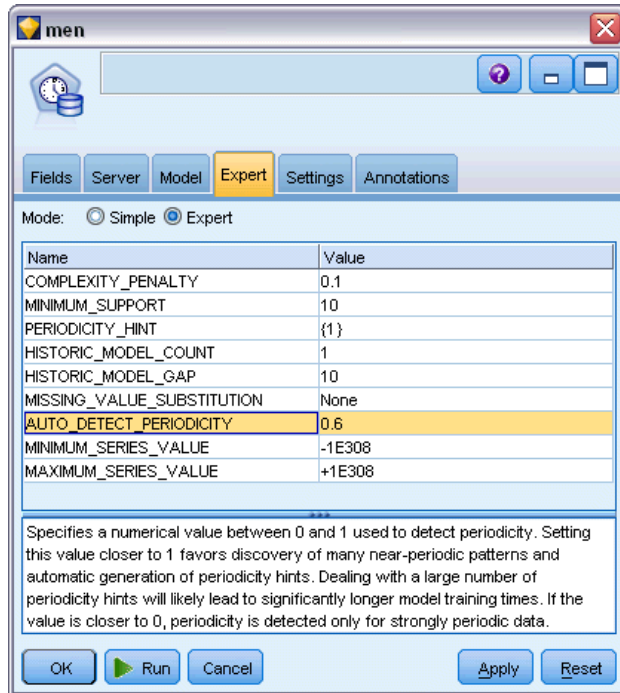
**Utiliser les données partitionnées.** Si un champ de partition est défini, cette option assure que seules les données provenant de la partition d'apprentissage sont utilisées pour générer le modèle. [Pour plus d'informations, reportez-vous à la section Noeud Partitionner dans le chapitre 4 dans Noeuds source, exécution et de sortie de IBM SPSS Modeler 15.](#)

**Avec extraction.** Si elle apparaît, cette option vous permet d'interroger le modèle pour obtenir des détails sur les observations incluses dans le modèle.

**Champ unique.** Dans la liste déroulante, sélectionnez le champ-clé temporel qui est utilisé pour créer un modèle de séries temporelles.

### Options de l'expert de séries temporelles MS

Figure 3-14  
Options de l'expert de séries temporelles MS

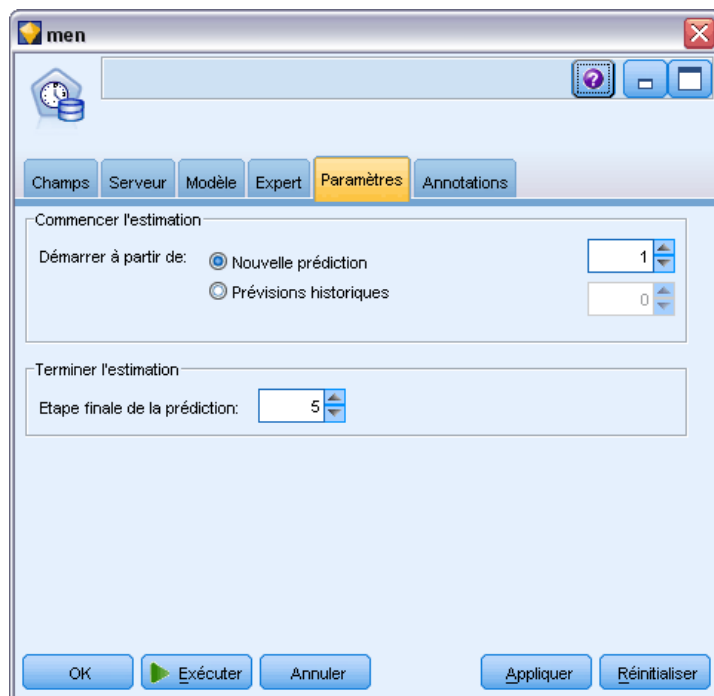


Les options disponibles dans l'onglet Expert peuvent varier en fonction de la structure du flux sélectionné. Consultez l'aide par champ de l'interface utilisateur pour obtenir des détails complets sur les options expert disponibles pour le noeud de modèle Analysis Services sélectionné.

Si vous effectuez des prévisions historiques, le nombre d'étapes historiques pouvant être incluses dans le résultat de scoring est décidé par la valeur de  $(\text{HISTORIC\_MODEL\_COUNT} * \text{HISTORIC\_MODEL\_GAP})$ . Par défaut, cette limite est de 10, ce qui signifie que 10 prévisions historiques uniquement seront effectuées. Dans ce cas, par exemple, une erreur se produit si vous saisissez une valeur inférieure à -10 pour Prévision historique dans l'onglet Paramètres du nugget de modèle (voir [Nugget de modèle de séries temporelles MS - Onglet Paramètres sur p. 45](#)). Si vous souhaitez afficher plus de prévisions historiques, vous pouvez augmenter la valeur de HISTORIC\_MODEL\_COUNT ou de HISTORIC\_MODEL\_GAP, mais cela augmentera la durée de création du modèle.

### Options des paramètres de séries temporelles MS

Figure 3-15  
Options des paramètres de séries temporelles MS



**Commencer l'estimation.** Spécifiez la période à laquelle les prédictions doivent commencer.

- **Commencer à partir de : Nouvelle prédiction.** La période à laquelle les prévisions futures doivent commencer, exprimée sous la forme d'un décalage par rapport à la dernière période de vos données historiques. Par exemple, si vos données historiques se sont terminées en 12/99 et que vous vouliez que vos prédictions commencent en 01/00, vous devez utiliser une valeur de 1 ; mais si vous souhaitez que vos prédictions commencent en 03/00, vous devez utiliser une valeur de 3.
- **Commencer à partir de : Prévission historique.** La période à laquelle les prévisions historiques doivent commencer, exprimée sous la forme d'un décalage négatif par rapport à la dernière période de vos données historiques. Par exemple, si vos données historiques se sont terminées en 12/99 et que vous souhaitez effectuer des prévisions historiques pour les cinq dernières périodes de vos données, vous devez utiliser une valeur de -5.

**Terminer l'estimation.** Spécifiez la période à laquelle les prédictions doivent se terminer.

- **Etape de fin de la prédiction.** La période à laquelle les prédictions se terminent, exprimée sous la forme d'un décalage par rapport à la dernière période de vos données historiques. Par exemple, si vos données historiques se terminent en 12/99 et que vous souhaitez que vos prédictions s'arrêtent en 06/00, vous devez utiliser la valeur 6. Pour les prédictions futures, la valeur doit toujours être supérieure ou égale à la valeur Commencer à partir de.

## **Noeud Classification de séquences MS**

Le noeud Classification de séquences MS utilise un algorithme d'analyse de séquences qui explore les données contenant des événements qui peuvent être liés en suivant des chemins ou *séquences*. On peut ainsi citer des exemples comme les chemins sur lesquels les utilisateurs cliquent lorsqu'ils naviguent ou recherchent un site Web ou l'ordre dans lequel un client ajoute des éléments à un panier chez un vendeur en ligne. L'algorithme recherche les séquences les plus courantes en regroupant, ou en *classifiant* les séquences identiques.

### **Conditions requises**

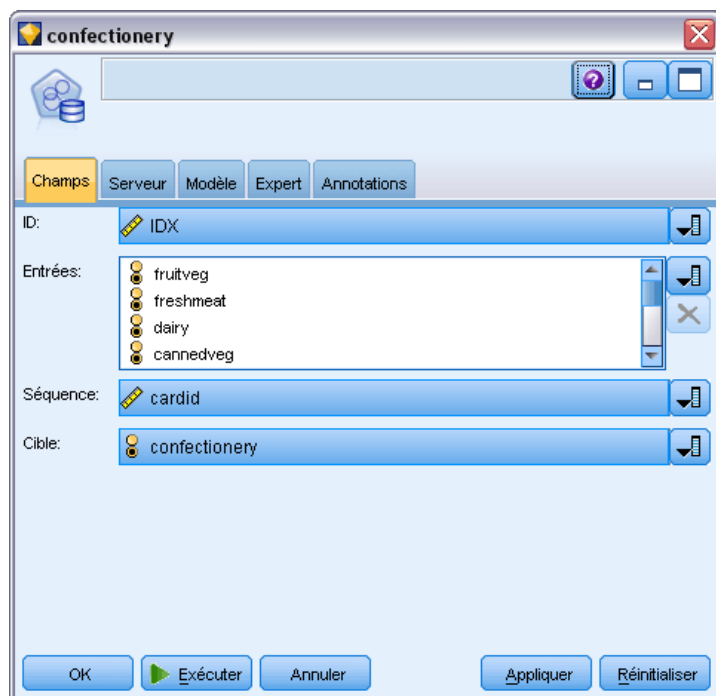
La configuration requise pour un modèle Classification de séquences Microsoft est :

- **Champ ID.** L'algorithme Classification de séquences Microsoft requiert que les informations de séquences soient stockées au format transactionnel (voir [Données tabulaires et données transactionnelles sur p.](#) ). Pour cela, un champ d'ID qui identifie chaque transaction est requis.
- **Au moins un champ d'entrée.** L'algorithme requiert au moins un champ d'entrée.
- **Champ de séquences.** L'algorithme nécessite également un champ d'identificateur de séquence qui doit avoir un niveau de mesure continu. Par exemple, vous pouvez utiliser un identificateur de page Web, un entier, ou une chaîne de texte, tant que le champ identifie les événements dans une séquence. Un seul identificateur de séquence est autorisé pour chaque séquence et un seul type de séquence est autorisé dans chaque modèle. Le champ Séquence doit être différent des champs ID et Unique.
- **Champ cible.** Un champ cible est requis lors de la création d'un modèle de classification de séquences.
- **Champ unique.** Un modèle de classification de séquences requiert un champ-clé qui identifie de manière unique les enregistrements. Vous pouvez définir le champ Unique pour qu'il soit le même que le champ ID.

### **Classification de séquences MS- Options de champs**

Tous les noeuds de modélisation comportent un onglet Champs, vous permettant de spécifier les champs à utiliser lors de la construction du modèle.

Figure 3-16  
Spécification des champs pour la classification de séquences MS



Avant de construire un modèle de classification de séquences, vous devez indiquer les champs à utiliser en tant que cibles et en tant qu'entrées. Veuillez noter que pour le noeud Classification de séquences MS, vous ne pouvez pas utiliser les informations de champs d'un noeud Typier en amont ; vous devez spécifier les paramètres de champ ici.

**ID.** Sélectionnez un champ ID dans la liste. Vous pouvez utiliser un champ numérique ou symbolique en tant que champ ID. Chaque valeur unique de ce champ doit indiquer une unité d'analyse spécifique. Par exemple, dans une analyse de paniers du marché, chaque ID peut représenter un client. Dans une analyse de l'utilisation du Web, chaque ID peut représenter un ordinateur (adresse IP) ou un utilisateur (ID de connexion).

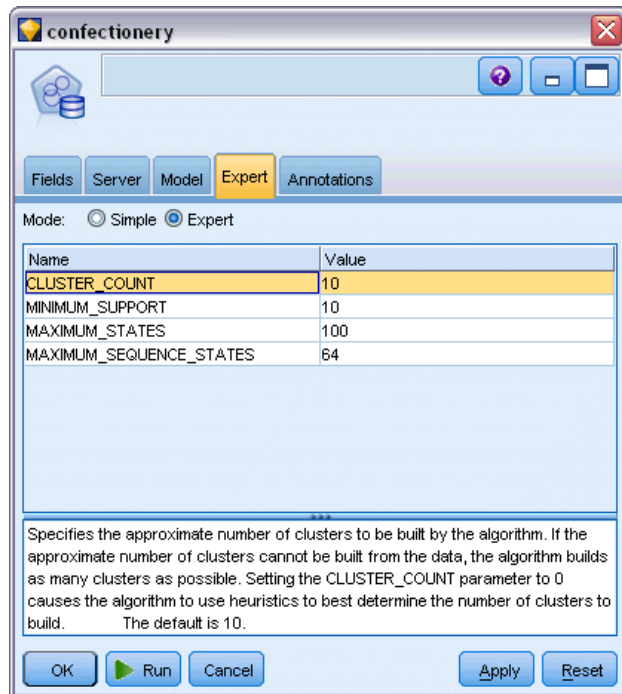
**Entrées.** Sélectionnez le ou les champs d'entrée du modèle. Ce sont les champs qui contiennent les événements d'intérêt concernant la modélisation des séquences.

**Séquence.** Choisissez un champ dans la liste que vous utiliserez comme champ d'identificateur de séquences. Par exemple, vous pouvez utiliser un identificateur de page Web, un entier, ou une chaîne de texte, tant que le champ identifie les événements dans une séquence. Un seul identificateur de séquence est autorisé pour chaque séquence et un seul type de séquence est autorisé dans chaque modèle. Le champ Séquence doit être différent du champ ID (spécifié sur cet onglet) et du champ Unique (spécifié sur l'onglet Modèle).

**Cible.** Choisissez un champ à utiliser comme champ cible c'est-à-dire le champ dont vous essayez de prédire la valeur en fonction des données de séquence.

### Options expert de la classification de séquences MS

Figure 3-17  
Spécification des options expert de la classification de séquences MS



Les options disponibles dans l'onglet Expert peuvent varier en fonction de la structure du flux sélectionné. Consultez l'aide par champ de l'interface utilisateur pour obtenir des détails complets sur les options expert disponibles pour le noeud de modèle Analysis Services sélectionné.

## Scoring des modèles Analysis Services

Le scoring des modèles est réalisé par Analysis Services dans SQL Server. Vous devrez peut-être envoyer l'ensemble de données à une table temporaire si les données proviennent de IBM® SPSS® Modeler ou doivent être préparées dans SPSS Modeler. Les modèles que vous créez depuis SPSS Modeler, via la fonction d'exploration de base de données, constituent en réalité un modèle distant géré sur le serveur de base de données ou de Data mining distant. Il s'agit là d'une distinction fondamentale que vous devez assimiler lorsque vous parcourez des modèles créés par le biais d'algorithmes Microsoft Analysis Services et que vous déterminez un score pour ces modèles.

Dans SPSS Modeler, en général, une seule prévision et une probabilité ou confiance associées sont livrées.

Pour voir des exemples d'évaluation de modèles, reportez-vous à [Exemples d'exploration Analysis Services sur p. 46](#).



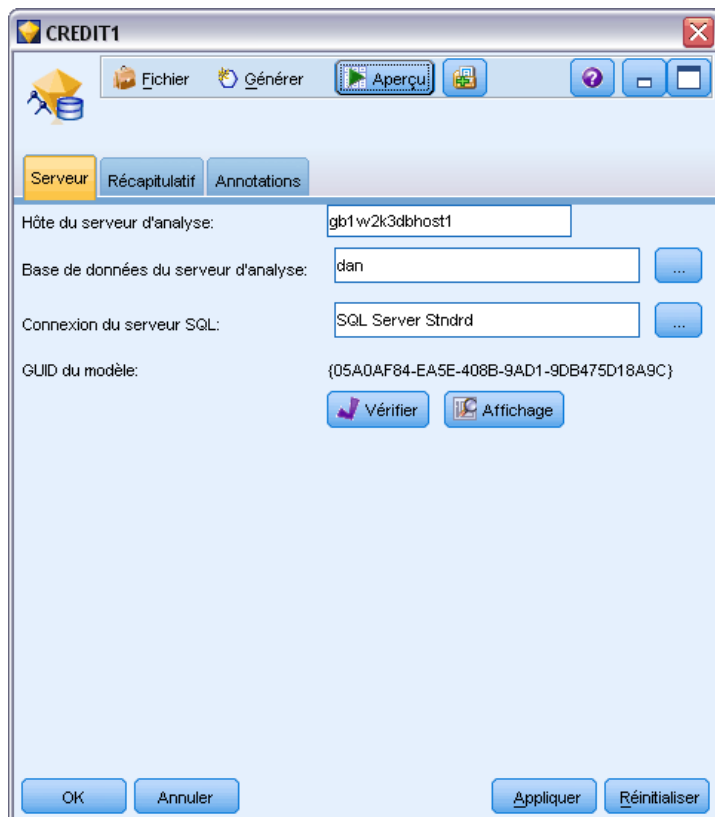
## Paramètres communs à tous les modèles Analysis Services

Les valeurs suivantes sont communes à tous les modèles Analysis Services.

### Onglet Serveur du nugget de modèle Analysis Services

L'onglet Serveur permet d'indiquer des connexions pour l'exploration d'une base de données. L'onglet fournit également la clé de modèle unique. Cette clé est générée de manière aléatoire lorsque le modèle est créé et est stocké dans le modèle dans IBM® SPSS® Modeler, et dans la description de l'objet de ce modèle stocké dans la base de données Analysis Services.

Figure 3-18  
Options serveur du nugget de modèle d'arbre décision MS

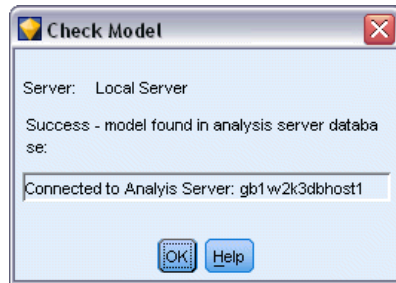


Dans l'onglet Serveur, vous pouvez configurer l'hôte et la base de données du serveur d'analyse, ainsi que la source de données SQL Server utilisée pour l'opération de scoring. Les options spécifiées dans cet onglet remplacent celles figurant dans les boîtes de dialogue des programmes externes ou de création de modèle dans IBM® SPSS® Modeler. [Pour plus d'informations, reportez-vous à la section Activation de l'intégration à Analysis Services sur p. 17.](#)

**GUID du modèle.** La clé du modèle est stipulée dans ce champ. Cette clé est générée de manière aléatoire lorsque le modèle est créé et est stocké dans le modèle dans SPSS Modeler, et dans la description de l'objet de ce modèle stocké dans la base de données Analysis Services.

**Vérifier.** Cliquez sur ce bouton pour vérifier la clé du modèle et celle du modèle stocké dans la base de données Analysis Services. Ainsi, vous pouvez vous assurer que le modèle figure toujours sur le serveur d'analyse. En outre, cela implique que la structure de ce modèle reste telle quelle.

Figure 3-19  
Résultats de la vérification des clés des modèles



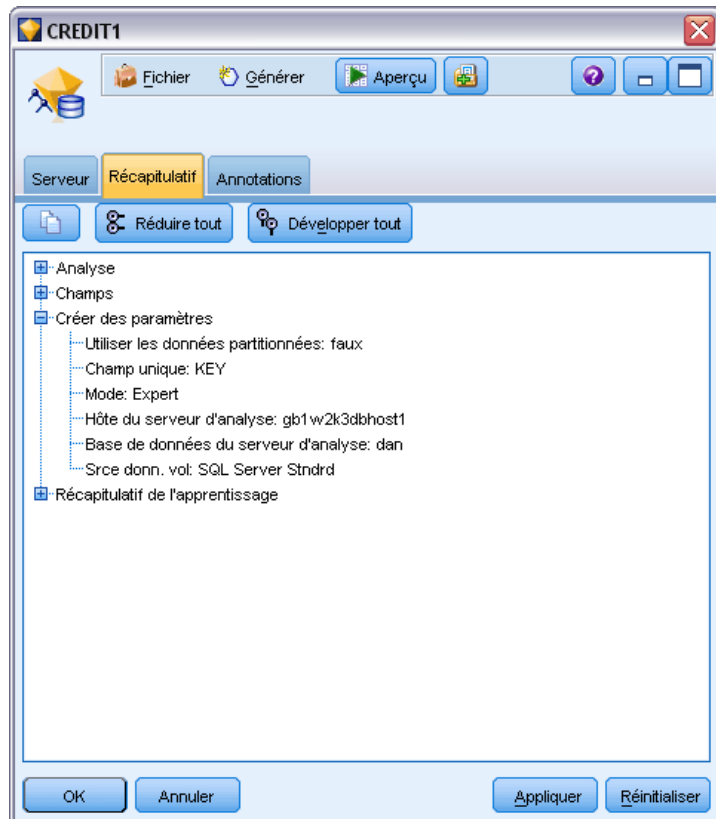
*Remarque :* le bouton Vérifier n'est disponible que pour les modèles ajoutés à l'espace de travail de flux lors de la préparation au scoring. Si la vérification échoue, cherchez à savoir si le modèle a été supprimé ou remplacé par un autre sur le serveur.

**Affichage.** Cliquez sur cette option pour afficher une vue graphique du modèle d'arbre décision. L'afficheur d'arbres décision est partagé par tous les algorithmes d'arbre décision dans SPSS Modeler. Les fonctions, quant à elles, sont identiques. [Pour plus d'informations, reportez-vous à la section Nuggets de modèle Arbre de décision dans le chapitre 6 dans \*Noeuds de modélisation de IBM SPSS Modeler 15\*.](#)

### Onglet Récapitulatif du nugget de modèle Analysis Services

Figure 3-20

Options Récapitulatif du nugget de modèle d'arbre décision MS



L'onglet Récapitulatif d'un nugget de modèle contient des informations sur le modèle lui-même (*Analyse*), sur les champs utilisés dans le modèle (*Champs*), sur les paramètres utilisés pour la construction du modèle (*Créer des paramètres*), ainsi que sur l'apprentissage du modèle (*Récapitulatif de l'apprentissage*).

Lorsque vous accédez au noeud pour la première fois, l'arborescence des résultats de l'onglet Récapitulatif est réduite. Pour afficher les résultats qui vous intéressent, utilisez la commande à gauche d'un élément pour le développer ou cliquez sur le bouton Développer tout pour afficher tous les résultats. Pour masquer les résultats lorsque vous avez terminé de les consulter, utilisez la commande de développement pour réduire les résultats voulus ou cliquez sur le bouton Réduire tout pour réduire tous les résultats.

**Analyse.** Affiche des informations sur le modèle en question. Si vous avez exécuté un noeud Analyse relié à ce nugget de modèle, les informations issues de l'analyse figureront également dans cette section. [Pour plus d'informations, reportez-vous à la section Noeud Analyse dans le chapitre 6 dans Noeuds source, exécution et de sortie de IBM SPSS Modeler 15.](#)

**Champs.** Répertoire les champs utilisés comme cibles et entrées lors de la création du modèle.

**Créer des paramètres.** Contient des informations sur les paramètres utilisés lors de la création du modèle.

**Récapitulatif de l'apprentissage.** Indique le type du modèle, le flux utilisé pour le créer, l'utilisateur qui l'a créé, ainsi que sa date et sa durée de création.

### ***Nugget du modèle Séries temporelles MS***

Le modèle de séries temporelles MS produit des scores uniquement pour les périodes prédites, pas pour les données historiques.

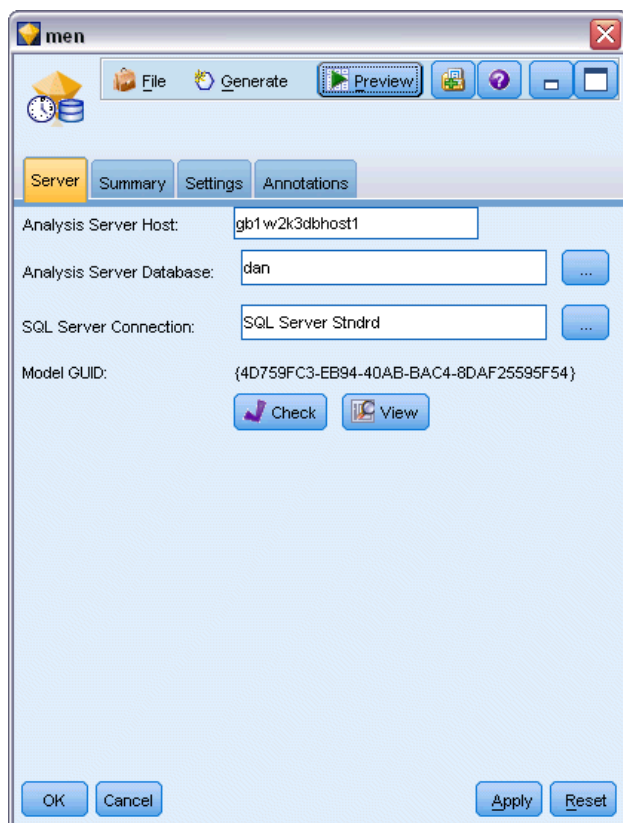
Les champs suivants sont ajoutés au modèle :

<b>Nom de champ</b>	<b>Description</b>
<i>\$M-champ</i>	Valeur prédite de <i>field</i>
<i>\$Var-champ</i>	Variance calculée de <i>champ</i>
<i>\$Stdev-champ</i>	Ecart-type de <i>champ</i>

### ***Nugget de modèle de séries temporelles MS - Onglet Serveur***

L'onglet Serveur permet d'indiquer des connexions pour l'exploration d'une base de données. L'onglet fournit également la clé de modèle unique. Cette clé est générée de manière aléatoire lorsque le modèle est créé et est stocké dans le modèle dans IBM® SPSS® Modeler, et dans la description de l'objet de ce modèle stocké dans la base de données Analysis Services.

Figure 3-21  
Options serveur du nugget de modèle de séries temporelles MS

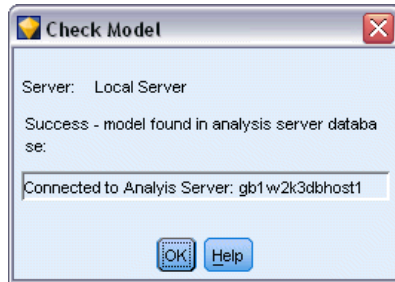


Dans l'onglet Serveur, vous pouvez configurer l'hôte et la base de données du serveur d'analyse, ainsi que la source de données SQL Server utilisée pour l'opération de scoring. Les options spécifiées dans cet onglet remplacent celles figurant dans les boîtes de dialogue des programmes externes ou de création de modèle dans IBM® SPSS® Modeler. [Pour plus d'informations, reportez-vous à la section Activation de l'intégration à Analysis Services sur p. 17.](#)

**GUID du modèle.** La clé du modèle est stipulée dans ce champ. Cette clé est générée de manière aléatoire lorsque le modèle est créé et est stocké dans le modèle dans SPSS Modeler, et dans la description de l'objet de ce modèle stocké dans la base de données Analysis Services.

**Vérifier.** Cliquez sur ce bouton pour vérifier la clé du modèle et celle du modèle stocké dans la base de données Analysis Services. Ainsi, vous pouvez vous assurer que le modèle figure toujours sur le serveur d'analyse. En outre, cela implique que la structure de ce modèle reste telle quelle.

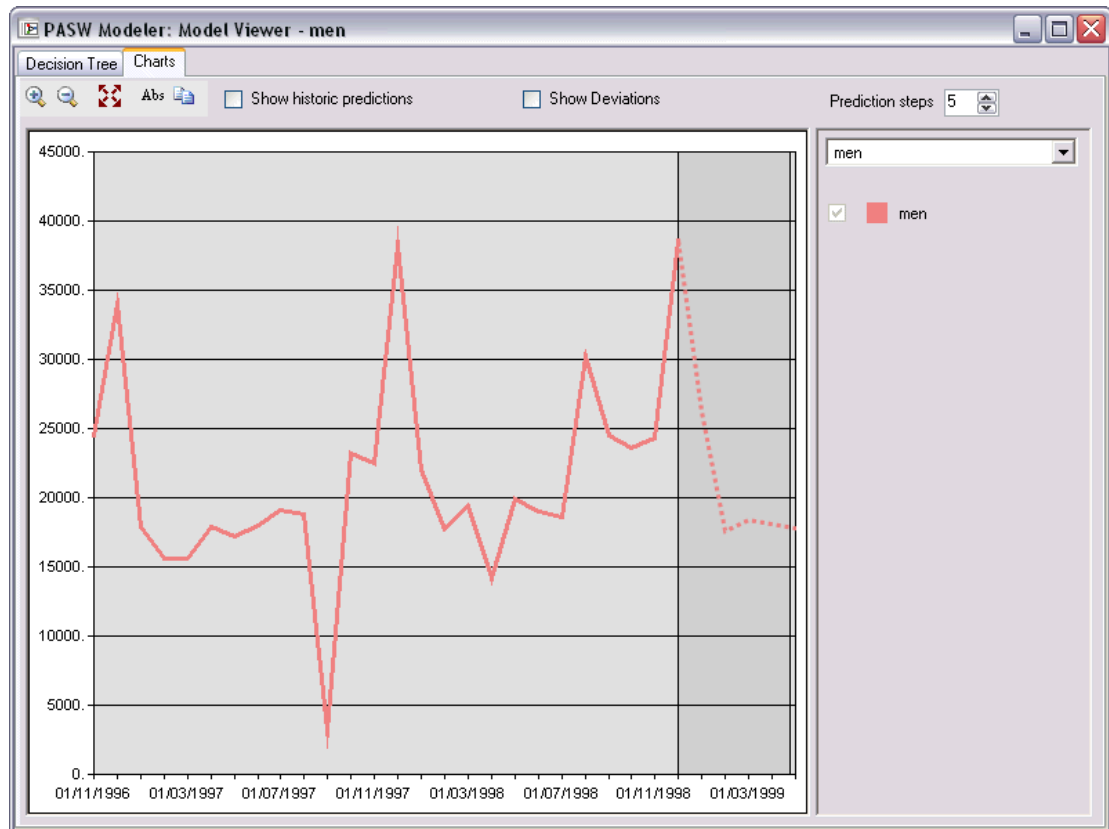
**Figure 3-22**  
Résultats de la vérification des clés des modèles



*Remarque* : le bouton Vérifier n'est disponible que pour les modèles ajoutés à l'espace de travail de flux lors de la préparation au scoring. Si la vérification échoue, cherchez à savoir si le modèle a été supprimé ou remplacé par un autre sur le serveur.

**Affichage.** Cliquez sur cette option pour afficher une vue graphique du modèle de séries temporelles. Analysis Services affiche le modèle terminé sous forme d'arbre. Vous pouvez également afficher un graphique qui affiche les valeurs historiques du champ cible sur une période de temps avec les valeurs prédites futures.

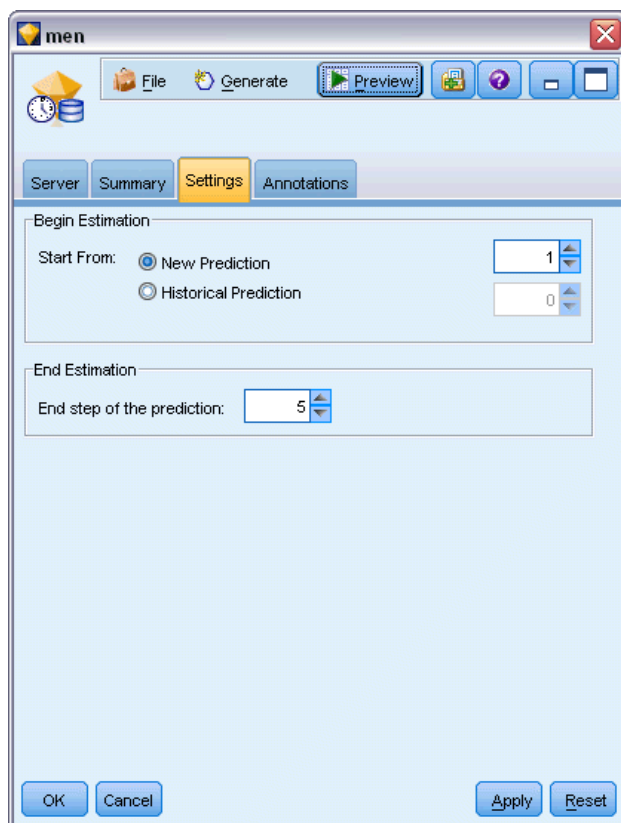
**Figure 3-23**  
Visualiseur de séries temporelles MS avec des valeurs historiques (ligne pleine) et des valeurs prédites futures (ligne en pointillés)



Pour plus d'informations, voir la description du visualiseur de séries temporelles dans la bibliothèque MSDN à l'adresse <http://msdn.microsoft.com/en-us/library/ms175331.aspx>.

### **Nugget de modèle de séries temporelles MS - Onglet Paramètres**

Figure 3-24  
Options paramètres du nugget de modèle de séries temporelles MS



**Commencer l'estimation.** Spécifiez la période à laquelle les prédictions doivent commencer.

- **Commencer à partir de : Nouvelle prédiction.** La période à laquelle les prévisions futures doivent commencer, exprimée sous la forme d'un décalage par rapport à la dernière période de vos données historiques. Par exemple, si vos données historiques se sont terminées en 12/99 et que vous voulez que vos prédictions commencent en 01/00, vous devez utiliser une valeur de 1 ; mais si vous souhaitez que vos prédictions commencent en 03/00, vous devez utiliser une valeur de 3.
- **Commencer à partir de : Prévission historique.** La période à laquelle les prévisions historiques doivent commencer, exprimée sous la forme d'un décalage négatif par rapport à la dernière période de vos données historiques. Par exemple, si vos données historiques se sont terminées en 12/99 et que vous souhaitez effectuer des prévisions historiques pour les cinq dernières périodes de vos données, vous devez utiliser une valeur de -5.

**Terminer l'estimation.** Spécifiez la période à laquelle les prédictions doivent se terminer.

- **Etape de fin de la prédiction.** La période à laquelle les prédictions se terminent, exprimée sous la forme d'un décalage par rapport à la dernière période de vos données historiques. Par exemple, si vos données historiques se terminent en 12/99 et que vous souhaitez que vos prédictions s'arrêtent en 06/00, vous devez utiliser la valeur 6. Pour les prédictions futures, la valeur doit toujours être supérieure ou égale à la valeur Commencer à partir de.

### ***Nugget de modèle de classification de séquences MS***

Les champs suivants sont ajoutés au modèle de classification de séquences MS (où *champ* est le nom du champ cible) :

Nom de champ	Description
<i>champ</i> -\$MC	Prédiction de la classe à laquelle cette séquence appartient.
<i>champ</i> -\$MCP	Probabilité que cette séquence appartienne à la classe prédite.
<i>champ</i> -\$MS	Valeur prédite de <i>field</i>
<i>champ</i> -\$MSP	Probabilité que la valeur du <i>champ</i> -\$MS soit correcte.

### ***Exportation de modèles et génération de noeuds***

Vous pouvez exporter un récapitulatif et une structure de modèle dans des fichiers au format texte et HTML. Vous pouvez générer les noeuds Sélectionner et Filtrer nécessaires. [Pour plus d'informations, reportez-vous à la section Navigation dans les nuggets de modèle dans le chapitre 3 dans \*Noeuds de modélisation de IBM SPSS Modeler 15\*.](#)

Semblables aux autres nuggets de modèles dans IBM® SPSS® Modeler, les nuggets de modèles Microsoft Analysis Services prennent en charge la création directe de noeuds d'opérations sur les enregistrements et les champs. A l'aide des options du menu Générer du nugget de modèle, vous pouvez créer les noeuds suivants :

- Noeud Sélectionner (uniquement si un élément est sélectionné dans l'onglet Modèle)
- Noeud Filtrer

### ***Exemples d'exploration Analysis Services***

Plusieurs exemples de flux sont fournis pour expliquer l'utilisation du Data mining MS Analysis Services avec IBM® SPSS® Modeler. Ces flux sont disponibles dans le dossier d'installation de SPSS Modeler à l'emplacement suivant :

`\Demos\Database_Modelling\Microsoft`

*Remarque* : Le dossier Démonstrations est accessible à partir du groupe de programmes IBM SPSS Modeler du menu Démarrer de Windows.



## Exemples de flux : Arbres décision

Les flux suivants peuvent être utilisés en séquence comme exemple du processus d'exploration de base de données via l'algorithme des arbres décision fourni par MS Analysis Services.

Flux Clementine	Description
<i>1_upload_data.str</i>	Utilisé pour nettoyer et envoyer des données à partir d'un fichier plat vers la base de données.
<i>2_explore_data.str</i>	Offre un exemple d'exploration des données à l'aide de IBM® SPSS® Modeler
<i>3_build_model.str</i>	Crée le modèle à l'aide de l'algorithme natif de base de données.
<i>4_evaluate_model.str</i>	Utilisé comme exemple d'évaluation de modèle avec SPSS Modeler
<i>5_deploy_model.str</i>	Permet de déployer le modèle de détermination des scores dans la base de données.

*Remarque* : Pour le bon déroulement de l'exemple, vous devez exécuter les flux dans l'ordre. En outre, vous devez mettre à jour les noeuds source et de modélisation de chaque flux de manière à référencer une source de données correcte pour la base de données à utiliser.

L'ensemble de données utilisé dans les exemples de flux concerne les applications pour carte de crédit et présente un problème de classification avec un mélange de valeurs prédites catégorielles et continues. Pour plus d'informations sur cet ensemble de données, reportez-vous au fichier *crx.names*, qui figure dans le dossier des exemples de flux.

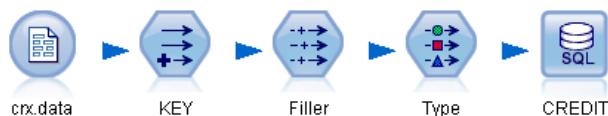
Cet ensemble de données est disponible à partir du référentiel d'apprentissage automatique UCI, sur le site suivant : <ftp://ftp.ics.uci.edu/pub/machine-learning-databases/credit-screening/>.

### Exemple de flux : Envoi de données

Le premier exemple de flux, *1\_upload\_data.str*, est utilisé pour nettoyer et envoyer les données d'un fichier plat dans SQL Server.

Figure 3-25

Exemple de flux utilisé pour envoyer les données



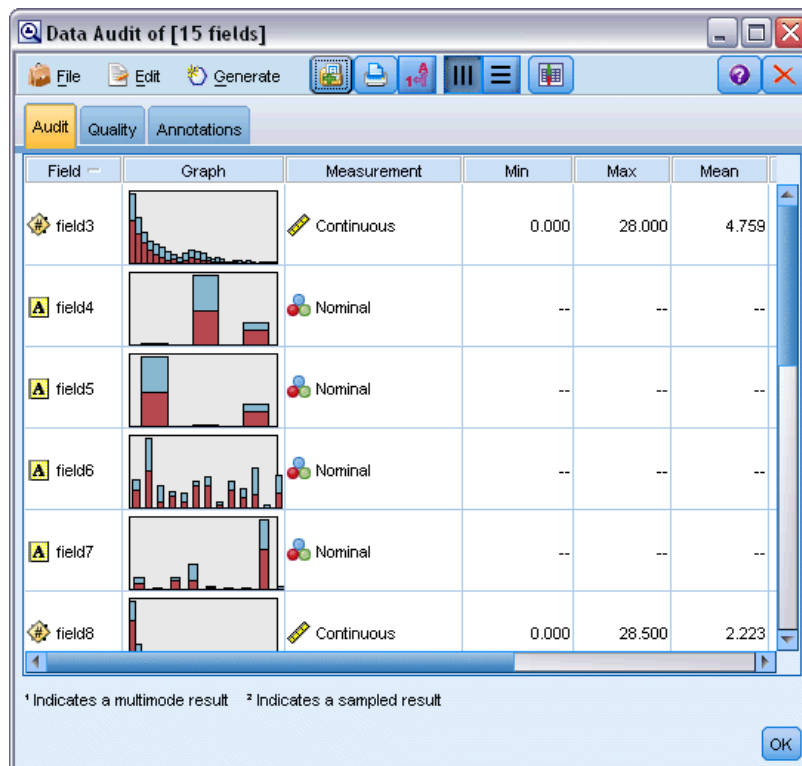
Le Data mining Analysis Services nécessitant un champ-clé, ce flux initial utilise un noeud Calculer pour ajouter un nouveau champ à l'ensemble de données *KEY*, doté des valeurs uniques 1,2,3, via la fonction IBM® SPSS® Modeler@INDEX.

Le noeud Remplacer suivant est utilisé pour le traitement des valeurs manquantes et remplace les champs vides lus dans le fichier texte *crx.data* par des valeurs *NULL*.

### Exemple de flux : Exploration des données

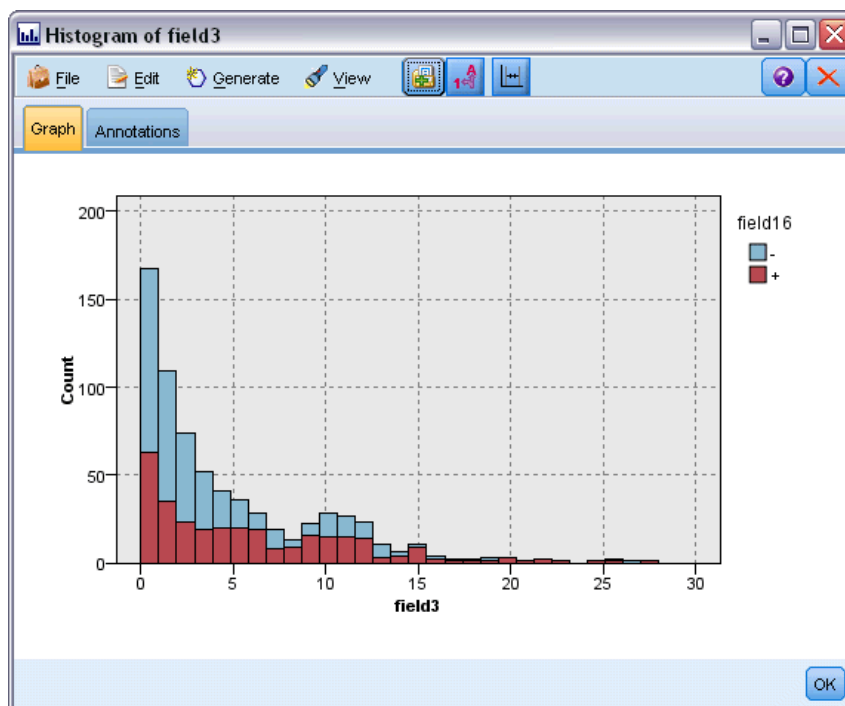
Le deuxième exemple de flux, *2\_explore\_data.str*, est utilisé pour décrire le noeud Audit données pour obtenir un aperçu général des données, y compris les statistiques récapitulatives et les graphiques. [Pour plus d'informations, reportez-vous à la section Noeud Audit données dans le chapitre 6 dans \*Noeuds source, exécution et de sortie de IBM SPSS Modeler 15\*.](#)

Figure 3-26  
Résultats de l'audit de données



Pour obtenir un graphique plus détaillé permettant d'explorer un champ de manière plus précise, double-cliquez sur le graphique voulu dans le rapport Audit données.

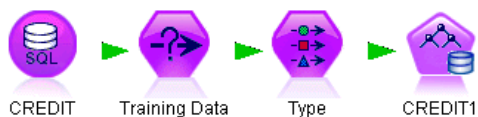
Figure 3-27  
Histogramme créé en double-cliquant sur un graphique dans la fenêtre Audit données



### Exemple de flux : Créer un modèle

Le troisième exemple de flux, *3\_build\_model.str*, illustre la création de modèles dans IBM® SPSS® Modeler. Vous pouvez relier le modèle de base de données au flux et double-cliquer dessus pour définir les paramètres de création.

Figure 3-28  
Exemple de flux de modélisation de base de données, où les noeuds en violet font référence à l'exécution d'opérations dans une base de données



Dans l'onglet Modèle de la boîte de dialogue, vous pouvez effectuer les tâches suivantes :

- ▶ Sélectionnez le champ d'ID unique Clé.

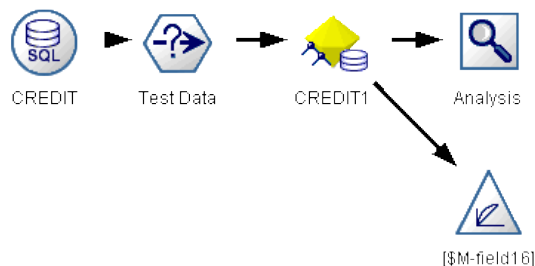
Dans l'onglet Expert, vous pouvez affiner les paramètres de création du modèle.

Avant l'exécution, veillez à indiquer la base de données correcte pour générer ce modèle. Utilisez l'onglet Serveur afin d'ajuster des paramètres.

### Exemple de flux : Evaluation d'un modèle

Le quatrième exemple de flux, `4_evaluate_model.str`, illustre les avantages que présente l'utilisation de IBM® SPSS® Modeler pour la modélisation dans la base de données. Une fois le modèle exécuté, vous pouvez l'ajouter à nouveau à votre flux de données et l'évaluer à l'aide des divers outils proposés par SPSS Modeler.

Figure 3-29  
Exemple de flux utilisé pour l'évaluation du modèle

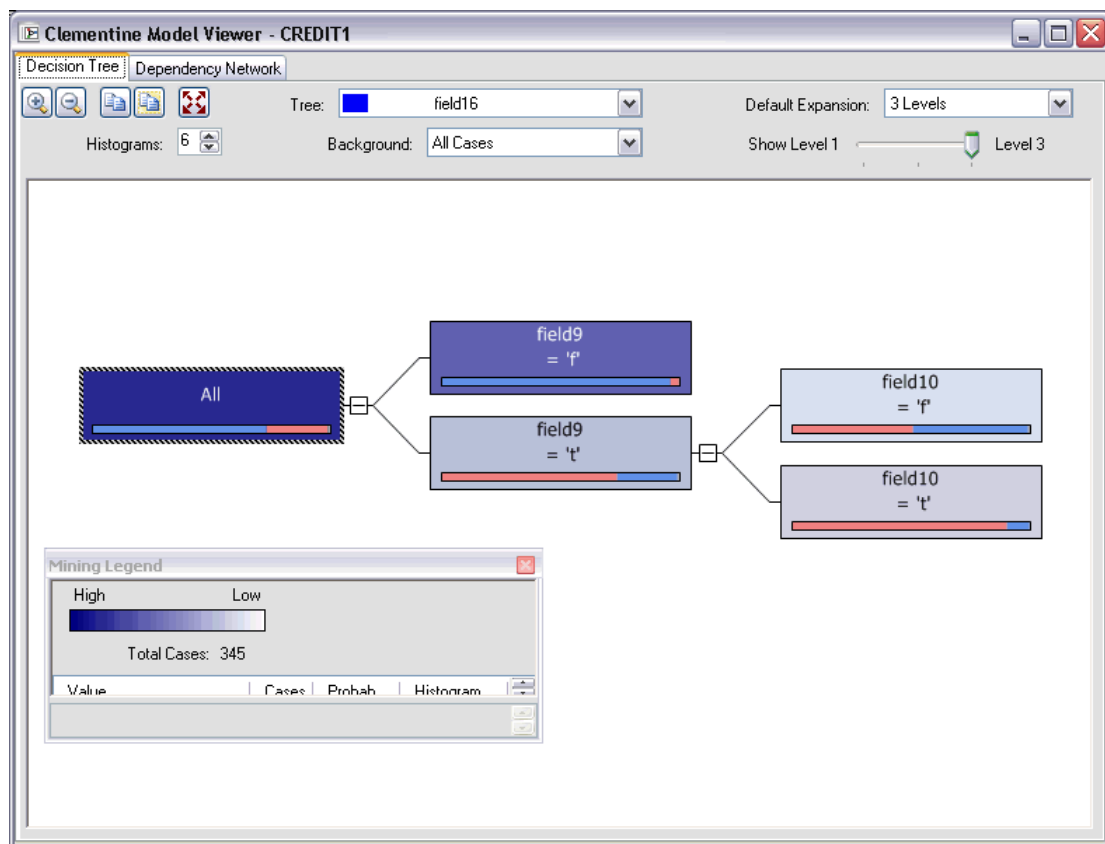


### Affichage des résultats d'une modélisation

Vous pouvez double-cliquer sur le nugget de modèle pour explorer vos résultats. L'onglet Récapitulatif affiche ces résultats en tant qu'arbre de règles. Vous pouvez également cliquer sur le bouton Affichage, dans l'onglet Serveur, pour obtenir une représentation graphique du modèle d'arbre décision.

Figure 3-30

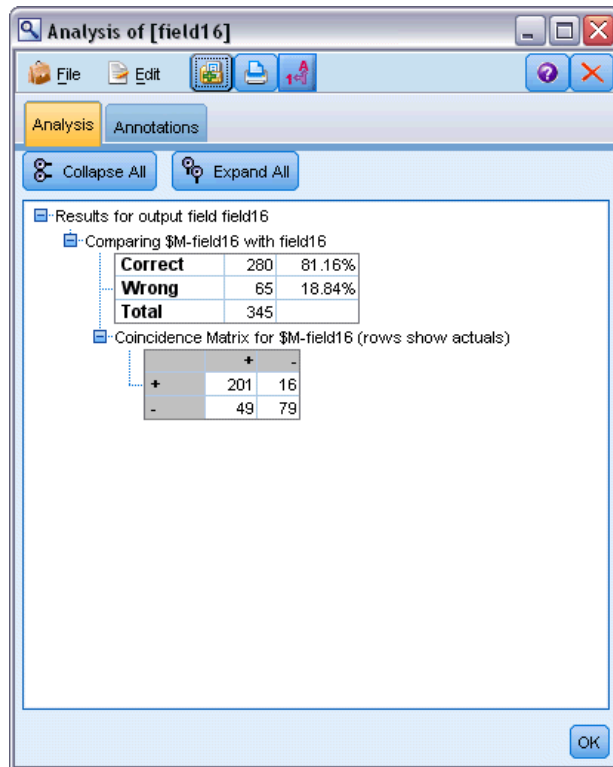
Visualiseur contenant le graphique des résultats du modèle d'arbre décision MS



### ***Evaluation des résultats d'un modèle***

Le noeud Analyse de l'exemple de flux crée une matrice de coïncidence illustrant le motif des correspondances entre chaque champ prédit et son champ cible. Exécutez le noeud Analyse pour afficher les résultats.

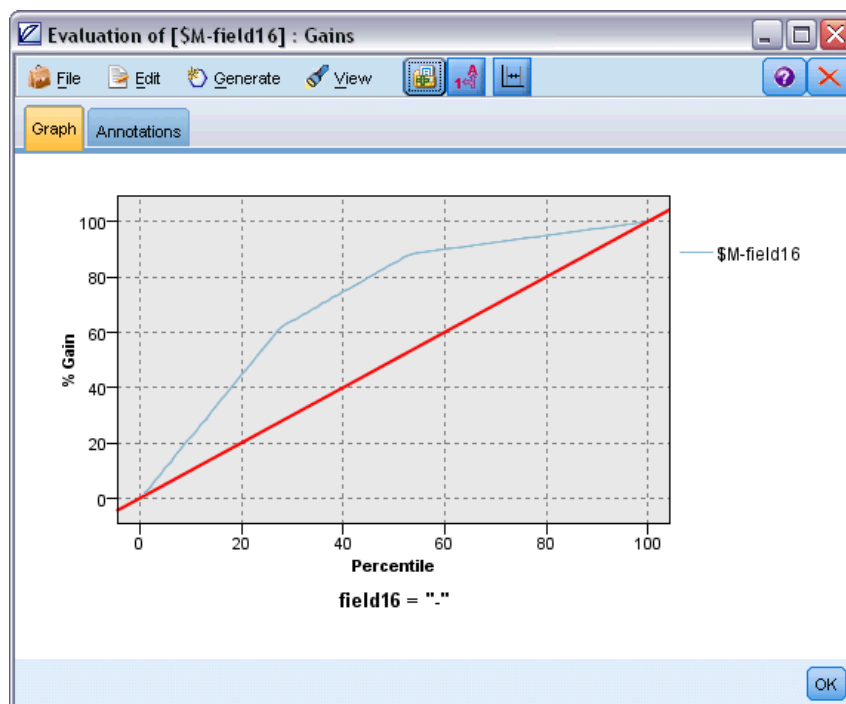
Figure 3-31  
Résultats du noeud Analyse



Le tableau indique que 81,16 % des prévisions générées par l’algorithme des arbres décision MS s’avèrent correctes.

Le noeud Evaluation de l’exemple de flux peut créer un graphique de gains qui affiche les améliorations de précision apportées par le modèle. Exécutez le noeud Evaluation pour afficher les résultats.

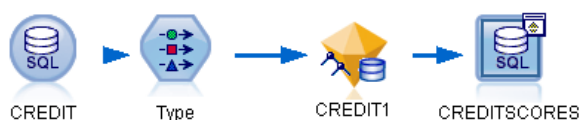
Figure 3-32  
Graphique Gains généré à l'aide du noeud Evaluation



### Exemple de flux : Déploiement d'un modèle

Une fois que la précision du modèle vous convient, vous pouvez le déployer pour l'utiliser avec des applications externes ou pour le republier dans la base de données. Dans le dernier exemple de flux, *5\_deploy\_model.str*, les données sont lues à partir de la table CREDIT, puis évaluées et publiées dans la table CREDITSCORES via un noeud Export SGBD.

Figure 3-33  
Exemple de flux utilisé pour déployer le modèle



L'exécution du flux génère l'instruction SQL suivante :

```
DROP TABLE CREDITSCORES
```

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

```

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

```



# ***Modélisation de bases de données à l'aide d'Oracle Data Mining***

## ***A propos d'Oracle Data Mining***

IBM® SPSS® Modeler prend en charge l'intégration d'Oracle Data Mining (ODM), offrant ainsi une famille d'algorithmes de Data mining étroitement incorporés dans le SGBDR Oracle. Ces fonctions sont accessibles via l'interface utilisateur graphique et l'environnement de développement orienté workflow de SPSS Modeler, ce qui permet aux clients d'utiliser les algorithmes de Data mining fournis par ODM.

SPSS Modeler prend en charge l'intégration des algorithmes suivants d'Oracle Data Mining :

- Naive Bayes
- Adaptive Bayes
- Support Vector Machine (SVM)
- Modèles linéaires généralisés (MLG)\*
- Arbre de décision
- O-Cluster
- k-Means
- Factorisation en matrices non négatives (NMF)
- A priori
- Description de longueur minimale (MDL)
- Importance de l'attribut (IA)

\* 11g R1 uniquement

## ***Conditions requises pour l'intégration à Oracle***

Les conditions préalables suivantes sont obligatoires pour réaliser une modélisation dans la base de données via Oracle Data Mining. N'hésitez pas à contacter l'administrateur de base de données pour vous assurer que ces conditions sont remplies.

- IBM® SPSS® Modeler en mode local ou en parallèle avec l'installation de IBM® SPSS® Modeler Server sous Windows ou sous UNIX.
- Oracle 10gR2 ou 11gR1 (base de données version 10.2 ou supérieure) doté de l'option Oracle Data Mining.

*Remarque* : 10gR2 offre une prise en charge de tous les algorithmes de modélisation de la base de données, à l'exception des Modèles linéaires généralisés (requiert 11gR1).

- Une source de données ODBC pour la connexion à Oracle, comme indiqué ci-après.

*Remarque* : La modélisation de base de données et l'optimisation SQL requièrent l'activation de la connectivité à SPSS Modeler Server sur l'ordinateur SPSS Modeler. Avec ce paramètre activé, vous pouvez accéder aux algorithmes de la base de données, effectuer le push back de SQL directement depuis SPSS Modeler et accéder à SPSS Modeler Server. Pour vérifier l'état actuel de la licence, choisissez ce qui suit dans le menu SPSS Modeler.

Aide > A propos de > Informations supplémentaires

Si la connectivité est activée, vous voyez l'option Activation du serveur dans l'onglet Etat de la licence.

Pour plus d'informations, reportez-vous à la section [Connexion au IBM SPSS Modeler Server](#) dans le chapitre 3 dans *Guide de l'utilisateur de IBM SPSS Modeler 15*.

## **Activation de l'intégration à Oracle**

Pour activer l'intégration de IBM® SPSS® Modeler à Oracle Data Mining, vous devez configurer Oracle et créer une source de données ODBC, activer cette intégration dans la boîte de dialogue Programmes externes de SPSS Modeler, puis activer la génération et l'optimisation SQL .

### **Configuration de Oracle**

Pour installer et configurer Oracle Data Mining, reportez-vous à la documentation Oracle, notamment au manuel *Oracle Administrator's Guide* pour plus d'informations.

### **Création d'une source de données ODBC pour Oracle**

Pour activer la connexion entre Oracle et SPSS Modeler, vous devez créer un nom de la source de données ODBC (DSN).

Avant de créer un DSN, vous devez avoir des connaissances de base des sources de données et des pilotes ODBC, ainsi que de la prise en charge de la base de données dans SPSS Modeler. [Pour plus d'informations, reportez-vous à la section Accès aux données dans le chapitre 2 dans Guide d'administration et des performances de IBM SPSS Modeler Server 15](#).

Si vous exécutez le système en mode réparti sur le IBM® SPSS® Modeler Server, créez le DSN sur l'ordinateur serveur. Si vous exécutez le système en mode (client) local, créez le DSN sur l'ordinateur client.

- ▶ Installez les pilotes ODBC. Ceux-ci sont disponibles sur le disque d'installation IBM® SPSS® Data Access Pack fourni avec cette version. Exécutez le fichier *setup.exe* pour démarrer le programme d'installation et sélectionnez les pilotes appropriés. Suivez les instructions à l'écran pour installer les pilotes.
- ▶ Créez le DSN (nom de source de données).

*Remarque* : La séquence de menus dépend de la version de Windows que vous utilisez.

- **Windows XP.** Dans le menu Démarrer, sélectionnez Panneau de configuration. Double-cliquez sur Outils d'administration, puis sur Sources de données (ODBC).

- **Windows Vista.** Dans le menu Démarrer, sélectionnez Panneau de configuration; puis Maintenance du système. Double-cliquez sur Outils d'administration, sélectionnez Sources de données (ODBC), puis cliquez sur Ouvrir.
- **Windows 7.** Dans le menu Démarrer, sélectionnez Panneau de configuration; puis Sécurité du système, puis Outils d'administration. Sélectionnez Sources de données (ODBC), puis cliquez sur Ouvrir.
- ▶ Cliquez sur l'onglet du DSN système, puis sur Ajouter.
- ▶ Sélectionnez le pilote SPSS OEM 6.0 Oracle Wire Protocol.
- ▶ Cliquez sur Terminer.
- ▶ Dans l'écran de configuration du pilote ODBC Oracle Wire Protocol, entrez le nom de la source de données de votre choix, le nom d'hôte du serveur Oracle, le numéro de port de la connexion, ainsi que le SID de l'instance Oracle utilisée.

Le nom d'hôte, le port et le SID figurent dans le fichier *tnsnames.ora* de l'ordinateur serveur si vous mettez en oeuvre TNS avec un fichier *tnsnames.ora*. Pour plus d'informations, contactez l'administrateur Oracle.

- ▶ Cliquez sur le bouton de test en vue de tester la connexion.

#### **Activation de l'intégration d'Oracle Data Mining dans IBM SPSS Modeler**

- ▶ Dans les menus SPSS Modeler, sélectionnez :  
Outils > Options > Programmes externes
- ▶ Cliquez sur l'onglet Oracle.

**Activer l'intégration d'Oracle Data Mining.** Active la palette Modélisation de la base de données (si elle n'est pas déjà visible) en bas de la fenêtre SPSS Modeler et ajoute les noeuds des algorithmes Oracle Data Mining.

**Connexion Oracle.** Indiquez la source de données ODBC Oracle par défaut permettant de créer et de stocker des modèles, ainsi qu'un nom d'utilisateur et un mot de passe valides. Ce paramètre peut être ignoré pour les noeuds de modélisation individuels et les nuggets de modèle.

*Remarque :* La connexion de base de données utilisée pour la modélisation peut être identique à celle utilisée pour accéder aux données. Par exemple, vous pouvez utiliser un flux qui accède aux données d'une base de données Oracle, les envoie vers SPSS Modeler pour qu'elles soient nettoyées ou fassent l'objet de diverses manipulations, puis les envoie à une autre base de données Oracle pour la modélisation. Les données d'origine peuvent également se trouver dans un fichier plat ou une autre source (non-Oracle), auquel cas elles doivent être envoyées à Oracle en vue de la modélisation. Dans tous les cas, les données sont automatiquement envoyées à une table temporaire créée dans la base de données utilisée pour la modélisation.

**Avertir au moment de remplacer un modèle Oracle Data Mining.** Sélectionnez cette option pour vous assurer que les modèles stockés dans la base de données ne sont pas ignorés par SPSS Modeler sans que vous en soyez informé.

**Répertoire des modèles Oracle Data Mining.** Affiche les modèles de Data mining disponibles.

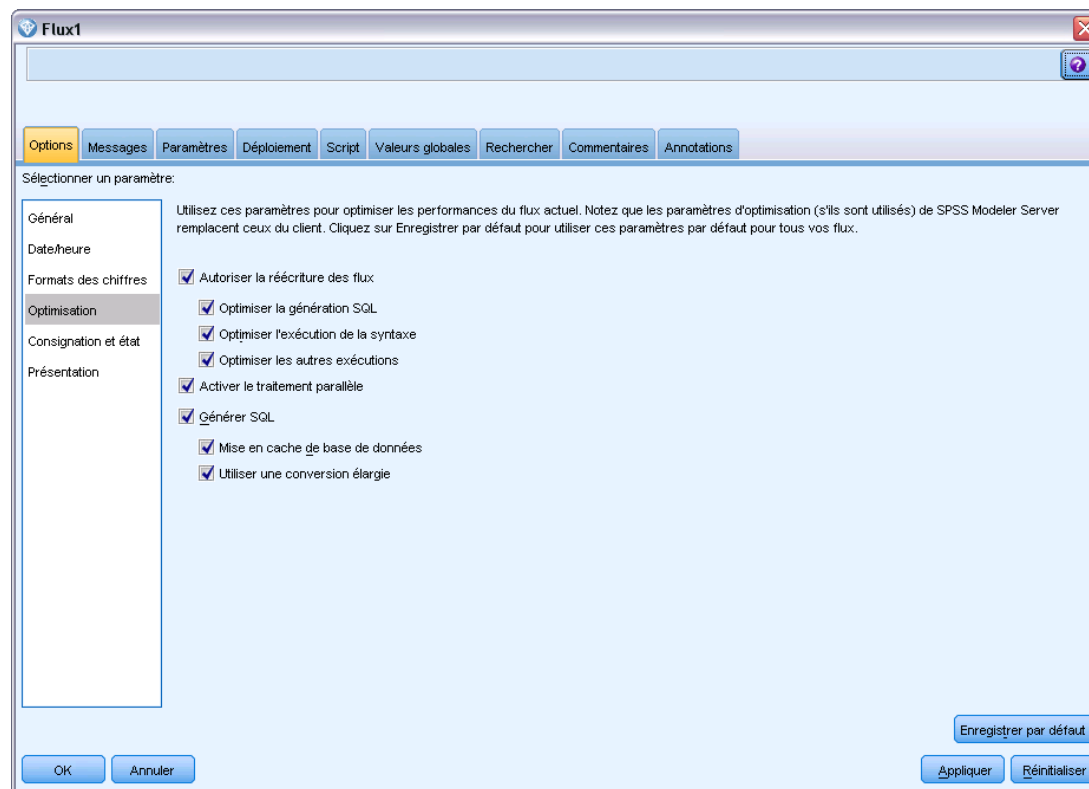
**Activer le lancement d'Oracle Data Miner. (facultatif)** Lorsque cette option est activée, SPSS Modeler lance l'application Oracle Data Miner. Pour plus d'informations, reportez-vous à [Oracle Data Miner sur p. 97](#).

**Chemin de l'exécutable d'Oracle Data Miner. (facultatif)** Indique l'emplacement physique du fichier exécutable d'Oracle Data Miner pour Windows (par exemple `C:\odm\bin\odminerw.exe`). Oracle Data Miner n'est pas installé avec SPSS Modeler ; la version appropriée doit être téléchargée sur le site Web d'Oracle (<http://www.oracle.com/technology/products/bi/odm/odminer.html>) et être installée comme cliente.

### Activation de la génération SQL et de l'optimisation

- Dans les menus SPSS Modeler, sélectionnez :  
Outils > Propriétés du flux > Options

Figure 4-1  
Paramètres d'optimisation



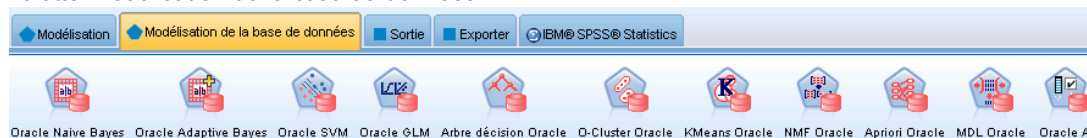
- Cliquez sur l'option Optimisation dans le volet de navigation.
- Confirmez que l'option Générer SQL est bien activée. Ce paramètre doit être utilisé pour que la modélisation de base de données puisse fonctionner.
- Sélectionnez Optimiser la génération SQL et Optimiser les autres exécutions (cette opération n'est pas forcément nécessaire, mais elle est vivement recommandée pour optimiser les performances).

Pour plus d'informations, reportez-vous à la section Définition des options d'optimisation pour les flux dans le chapitre 5 dans *Guide de l'utilisateur de IBM SPSS Modeler 15*.

## Création de modèles à l'aide d'Oracle Data Mining

Les noeuds de création de modèle Oracle présentent les mêmes principes de fonctionnement que les autres noeuds de modélisation IBM® SPSS® Modeler, à quelques exceptions près. Vous pouvez accéder à ces noeuds depuis la palette Modélisation de la base de données au bas de la fenêtre SPSS Modeler.

Figure 4-2  
Palette Modélisation de la base de données



### Remarques sur les données

Oracle exige que les données catégorielles soient stockées dans un format de chaîne (CHAR ou VARCHAR2). Par conséquent, SPSS Modeler n'autorise pas la spécification des champs de stockage numérique avec un niveau de mesure considéré comme *Booléen* ou *Nominal* (catégoriels) en tant que données d'entrée des modèles ODM. Si nécessaire, les valeurs peuvent être converties en chaînes dans SPSS Modeler à l'aide du noeud Recoder. [Pour plus d'informations, reportez-vous à la section Noeud Recoder dans le chapitre 4 dans Noeuds source, exécution et de sortie de IBM SPSS Modeler 15.](#)

**Champ cible.** Un seul champ peut être sélectionné en tant que champ (cible) de sortie dans les modèles de classification ODM.

**Nom du modèle.** Depuis la version Oracle 11gR1, le nom unique est un mot-clé et ne peut être utilisé comme nom de modèle personnalisé.

**Champ unique.** Indique le champ permettant d'identifier chaque observation de façon unique. Par exemple, il peut s'agir d'un champ d'ID, comme *ID client*. SPSS Modeler impose une restriction selon laquelle ce champ-clé doit être numérique.

*Remarque :* Ce champ est en option pour tous les noeuds Oracle à l'exception des noeuds Oracle Adaptive Bayes, Oracle O-Cluster et Oracle Apriori.

### Commentaires généraux

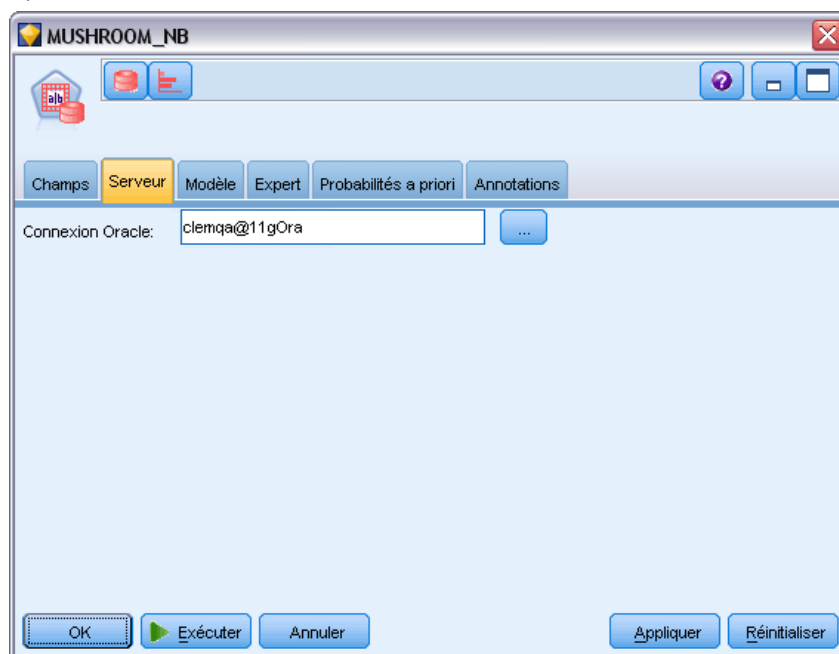
- La fonction d'exportation/importation PMML n'est pas disponible dans SPSS Modeler pour les modèles créés par Oracle Data Mining.
- Le scoring des modèles s'effectue systématiquement dans ODM. Vous devrez peut-être envoyer l'ensemble de données dans une table temporaire si les données proviennent ou doivent être préparées dans SPSS Modeler.
- Dans SPSS Modeler, en général, une seule prévision et une probabilité ou confiance associées sont livrées.

- SPSS Modeler limite à 1 000 le nombre de champs utilisés pour la création et le scoring de modèles.
- SPSS Modeler peut déterminer le score des modèles ODM à partir de flux publiés pour être exécutés via IBM® SPSS® Modeler Solution Publisher. [Pour plus d'informations, reportez-vous à la section Fonctionnement de IBM SPSS Modeler Solution Publisher dans le chapitre 2 dans IBM SPSS Modeler 15 Solution Publisher.](#)

### Options du serveur de modèles Oracle

Spécifiez la connexion Oracle permettant d'envoyer des données pour la modélisation. Si nécessaire, vous pouvez sélectionner une connexion dans l'onglet Serveur de chaque noeud de modélisation pour annuler la connexion Oracle par défaut indiquée dans la boîte de dialogue Programmes externes. [Pour plus d'informations, reportez-vous à la section Activation de l'intégration à Oracle sur p. 56.](#)

Figure 4-3  
Options du serveur Oracle



#### Commentaires

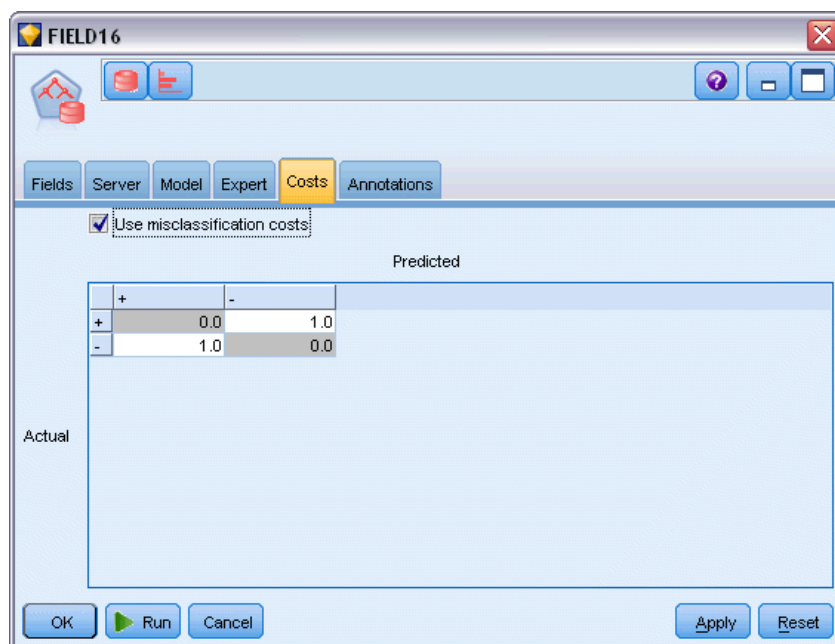
- La connexion utilisée pour la modélisation peut être identique à la connexion utilisée dans le noeud source d'un flux. Par exemple, vous pouvez utiliser un flux qui accède aux données d'une base de données Oracle, les envoie vers IBM® SPSS® Modeler pour qu'elles soient

nettoyées ou fassent l'objet de diverses manipulations, puis les envoie à une autre base de données Oracle pour la modélisation.

- Le nom de la source de données ODBC est incorporé de manière efficace dans chaque flux SPSS Modeler. Si un flux créé sur un hôte est exécuté sur un autre hôte, le nom de la source de données doit être identique sur chaque hôte. Vous pouvez également sélectionner une autre source de données dans l'onglet Serveur de chaque source ou noeud de modélisation.

## Coûts de mauvaise réaffectation

Figure 4-4  
Options des coûts Oracle



Selon le contexte, certains types d'erreur peuvent se révéler plus coûteux que d'autres. Par exemple, il peut être plus coûteux de classer un candidat au crédit à haut risque dans la catégorie à faible risque (un type d'erreur) que de classer un candidat à faible risque dans la catégorie à haut risque (un autre type d'erreur). L'option des coûts de mauvaise réaffectation vous permet de spécifier l'importance relative de différentes erreurs de prévision.

Les coûts d'une mauvaise réaffectation sont des pondérations appliquées à des revenus définis. Elles sont prises en compte dans le modèle et peuvent modifier la prévision (ce qui permet d'éviter des erreurs qui pourraient coûter cher).

A l'exception des modèles C5.0, les coûts d'une mauvaise réaffectation ne s'appliquent lorsque vous évaluez un modèle et ne sont pas pris en compte lors du classement ou de la comparaison de modèles par le biais d'un noeud Classificateur automatique, d'un graphique Evaluation ou d'un noeud Analyse. Il se peut qu'un modèle comprenant des coûts ne produise pas moins d'erreurs qu'un modèle n'en comprenant pas et ne classe pas de façon maximale en termes de précision générale. En revanche, il est probable, qu'en pratique, ses performances soient meilleures du fait qu'il dispose de biais intégrés rendant les erreurs *moins coûteuses*.

La matrice de mauvaise réaffectation des coûts affiche le coût de chaque combinaison possible de catégories prédites et de catégories réelles. Par défaut, tous les coûts de mauvaise réaffectation sont paramétrés sur 1. Pour entrer des valeurs de coût personnalisées, sélectionnez Utiliser les coûts de mauvaise réaffectation et entrez vos valeurs personnalisées dans la matrice des coûts.

Pour modifier un coût dû à une mauvaise réaffectation, sélectionnez la cellule correspondant à la combinaison voulue de valeurs prédites et de valeurs réelles, supprimez le contenu de la cellule et entrez le coût à appliquer à la cellule. Les coûts ne sont pas automatiquement symétriques. Ainsi, si vous définissez le coût d'une mauvaise affectation de  $A$  en tant que  $B$  sur 2, le coût d'une mauvaise affectation de  $B$  en tant que  $A$  sera toujours défini sur la valeur par défaut 1, à moins que vous ne modifiez cette valeur de manière explicite.

*Remarque* : Seul le modèle d'arbre décision permet d'indiquer les coûts au moment de la création.

## **Oracle Naive Bayes**

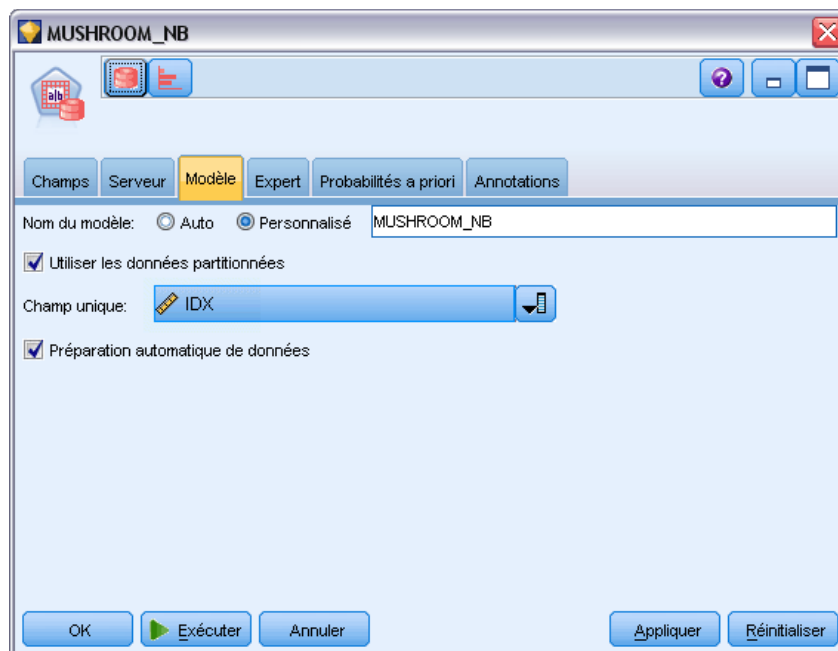
Naive Bayes est un algorithme bien connu pour les problèmes de classification. Ce modèle est appelé *naïve*, car il considère toutes les variables de prévision proposées comme étant indépendantes les unes des autres. Naive Bayes est un algorithme rapide et évolutif qui calcule les probabilités conditionnelles des combinaisons d'attributs et de l'attribut cible. Une probabilité indépendante est établie à partir des données d'apprentissage. Cette probabilité détermine la vraisemblance de chaque classe cible et est calculée en fonction de l'occurrence de chaque catégorie de valeur issue des variables d'entrée.

- La validation croisée permet de tester la précision d'un modèle sur les mêmes données que celles utilisées pour le créer. Elle s'avère particulièrement utile lorsque le nombre d'observations disponibles pour créer un modèle est réduit.
- Vous pouvez parcourir la sortie du modèle dans une matrice. Les valeurs figurant dans cette matrice constituent des probabilités conditionnelles liant les classes prédites (colonnes) aux combinaisons variable indépendante-valeur (lignes).



## Options des modèles Naive Bayes

Figure 4-5  
Options des modèles Naive Bayes



**Nom du modèle.** Vous pouvez générer le nom du modèle automatiquement sur la base du champ cible ou ID (ou du type de modèle si aucun de ces champs n'est spécifié) ou spécifier un nom personnalisé.

**Utiliser les données partitionnées.** Si un champ de partition est défini, cette option assure que seules les données provenant de la partition d'apprentissage sont utilisées pour générer le modèle. [Pour plus d'informations, reportez-vous à la section Noeud Partitionner dans le chapitre 4 dans Noeuds source, exécution et de sortie de IBM SPSS Modeler 15.](#)

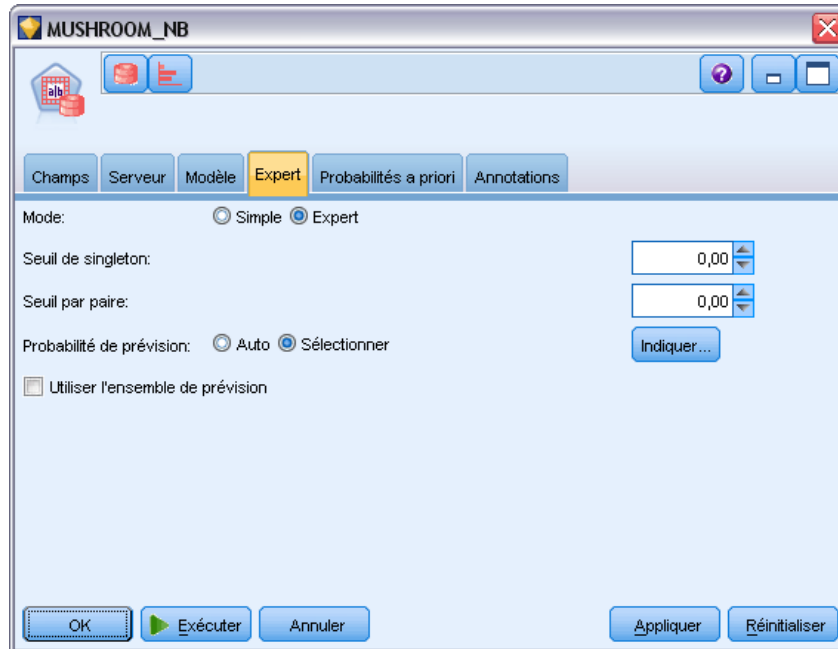
**Champ unique.** Indique le champ permettant d'identifier chaque observation de façon unique. Par exemple, il peut s'agir d'un champ d'ID, comme *ID client*. IBM® SPSS® Modeler impose une restriction selon laquelle ce champ-clé doit être numérique.

*Remarque :* Ce champ est en option pour tous les noeuds Oracle à l'exception des noeuds Oracle Adaptive Bayes, Oracle O-Cluster et Oracle Apriori.

**Préparation automatique de données.** (11g uniquement) Active (par défaut) ou désactive le mode de préparation automatique de données d'Oracle Data Mining. Si cette case est cochée, ODM effectue automatiquement les transformations de données nécessaires à l'algorithme. Pour des informations supplémentaires, consultez le manuel *Oracle Data Mining Concepts*.

## Options expert Naive Bayes

Figure 4-6  
Options expert Naive Bayes



Lorsque le modèle est créé, les valeurs d'attribut de variable indépendante ou les paires de valeurs ne sont pas prises en compte, sauf si les occurrences d'une valeur ou d'une paire de valeurs donnée sont en nombre suffisant dans les données d'apprentissage. Les seuils de non-prise en compte des valeurs sont fournis en tant que fractions basées sur le nombre d'enregistrements de ces données d'apprentissage. Si vous ajustez ces seuils, vous pouvez réduire le nombre de parasites et rendre le modèle plus à même de s'étendre à d'autres ensembles de données.

- **Seuil de singleton.** Indique le seuil d'une valeur d'attribut de variable indépendante donnée. Le nombre d'occurrences d'une valeur doit être égal ou supérieur à la fraction fournie ou cette valeur n'est pas prise en compte.
- **Seuil par paire.** Indique le seuil d'une paire de valeurs de variable indépendante et d'attribut donnée. Le nombre d'occurrences d'une paire de valeurs doit être égal ou supérieur à la fraction fournie ou cette paire n'est pas prise en compte.

**Probabilité de prévision.** Permet au modèle d'inclure la probabilité d'une prévision correcte pour un résultat possible du champ cible. Pour activer cette fonction, sélectionnez Sélectionner, cliquez sur le bouton Spécifier, choisissez l'un des résultats possibles et cliquez sur Insérer.

**Utiliser l'ensemble de prévision.** Génère un tableau des résultats possibles du champ cible.

## Oracle Adaptive Bayes

Adaptive Bayes Network (ABN) crée des classificateurs réseau bayésiens via la méthode MDL (Minimum Description Length) et la sélection de fonction automatique. Les résultats ABN sont parfois meilleurs que les résultats Naive Bayes, et tout au moins équivalents, même si les

performances en sont ralenties. L'algorithme ABN permet de créer trois types de modèle bayésien avancé, y compris les modèles d'arbre décision simplifiés (à fonction unique), les modèles Naive Bayes élagués et les modèles multifonctions améliorés.

### **Modèles générés**

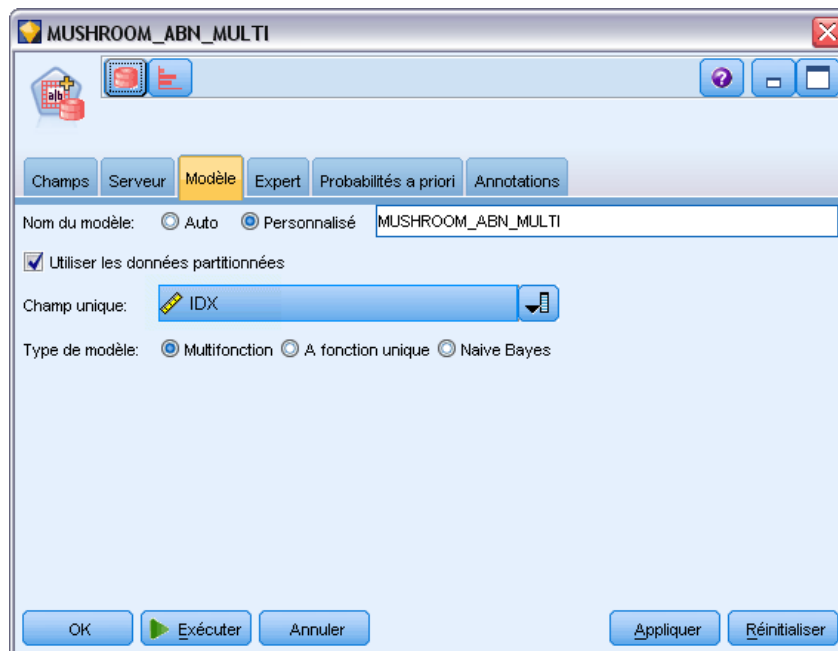
En mode de création à fonction unique, ABN génère un arbre décision simplifié, à partir d'un ensemble de règles lisibles, qui permet à l'utilisateur ou à l'analyste de comprendre les données de base des prévisions du modèle, et d'agir ou de fournir des explications aux autres utilisateurs en conséquence. Cela peut constituer un avantage significatif par rapport aux modèles Naive Bayes et aux modèles multifonctions. Vous pouvez parcourir ces règles comme un ensemble de règles standard dans IBM® SPSS® Modeler. Un ensemble de règles simple peut avoir la syntaxe suivante :

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

Vous ne pouvez pas parcourir les modèles Naive Bayes élagués ni les modèles multifonctions dans SPSS Modeler.

## **Options des modèles Adaptive Bayes**

Figure 4-7  
Options des modèles Adaptive Bayes



**Nom du modèle.** Vous pouvez générer le nom du modèle automatiquement sur la base du champ cible ou ID (ou du type de modèle si aucun de ces champs n'est spécifié) ou spécifier un nom personnalisé.

**Utiliser les données partitionnées.** Si un champ de partition est défini, cette option assure que seules les données provenant de la partition d'apprentissage sont utilisées pour générer le modèle. [Pour plus d'informations, reportez-vous à la section Noeud Partitionner dans le chapitre 4 dans \*Noeuds source, exécution et de sortie de IBM SPSS Modeler 15\*.](#)

**Champ unique.** Indique le champ permettant d'identifier chaque observation de façon unique. Par exemple, il peut s'agir d'un champ d'ID, comme *ID client*. IBM® SPSS® Modeler impose une restriction selon laquelle ce champ-clé doit être numérique.

*Remarque :* Ce champ est en option pour tous les noeuds Oracle à l'exception des noeuds Oracle Adaptive Bayes, Oracle O-Cluster et Oracle Apriori.

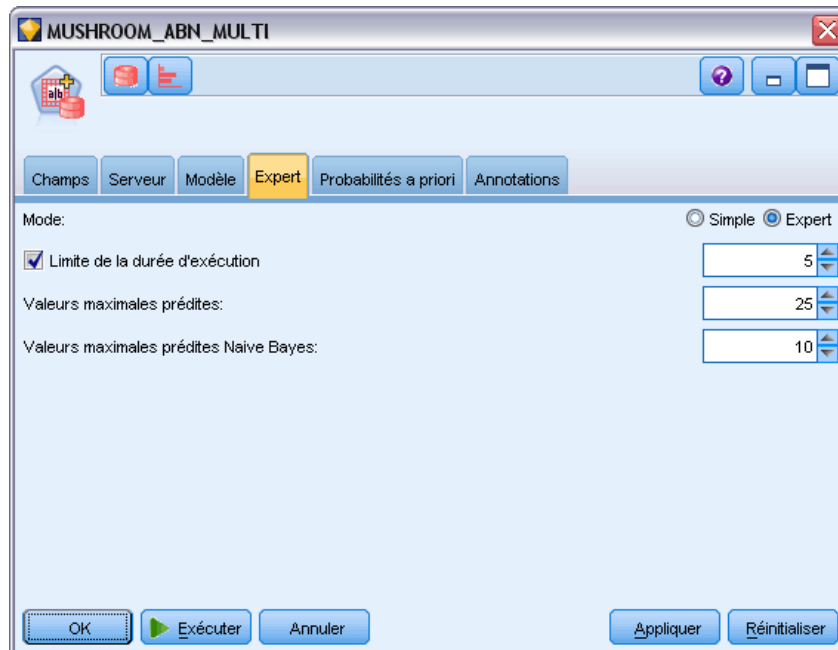
### **Type de modèle**

Vous avez le choix entre trois modes pour créer un modèle.

- **Multifonction.** Crée et compare plusieurs modèles, y compris un modèle NB et des modèles de probabilité de produit à fonction unique et multifonctions. Ce mode est le plus complet. Par conséquent, les calculs effectués sont généralement plus longs. Des règles sont créées seulement si le modèle à fonction unique s'avère être le modèle optimal. Si un modèle multifonction ou NB est sélectionné, aucune règle n'est générée.
- **A fonction unique.** Crée un arbre décision simplifié basé sur un ensemble de règles. Chaque règle contient une condition, ainsi que des probabilités associées à chaque sortie. Les règles sont mutuellement exclusives et sont fournies dans un format lisible, ce qui peut constituer un avantage significatif par rapport aux modèles Naive Bayes et aux modèles multifonctions.
- **Naive Bayes.** Crée un modèle NB simple et le compare à la probabilité a priori de l'échantillon global (proportion des valeurs cible de cet échantillon). Le modèle NB n'est généré comme sortie que s'il s'avère mieux prédire les valeurs cible que cette probabilité a priori. Dans le cas contraire, aucun modèle n'est généré comme sortie.

## Options expert Adaptive Bayes

Figure 4-8  
Options expert Adaptive Bayes



**Limite de la durée d'exécution.** Sélectionnez cette option pour indiquer la durée maximale de création, en minutes. Vous pouvez ainsi écourter le temps de génération des modèles, bien qu'ils risquent d'être moins précis au final. A chaque étape du processus de modélisation, l'algorithme vérifie s'il peut passer à l'étape suivante dans le délai spécifié avant de poursuivre, puis renvoie le modèle optimal disponible une fois le laps de temps maximal écoulé.

**Valeurs maximales prédites.** Cette option permet de limiter la complexité du modèle et d'améliorer les performances en réduisant le nombre de variables indépendantes utilisé. Les variables indépendantes sont classées en fonction de la mesure MDL de la corrélation à la cible utilisée comme mesure de la probabilité selon laquelle ils peuvent être ajoutés au modèle.

**Valeurs maximales prédites Naive Bayes.** Cette option indique le nombre maximal de variables indépendantes à utiliser dans le modèle Naive Bayes.

## Oracle Support Vector Machine (SVM)

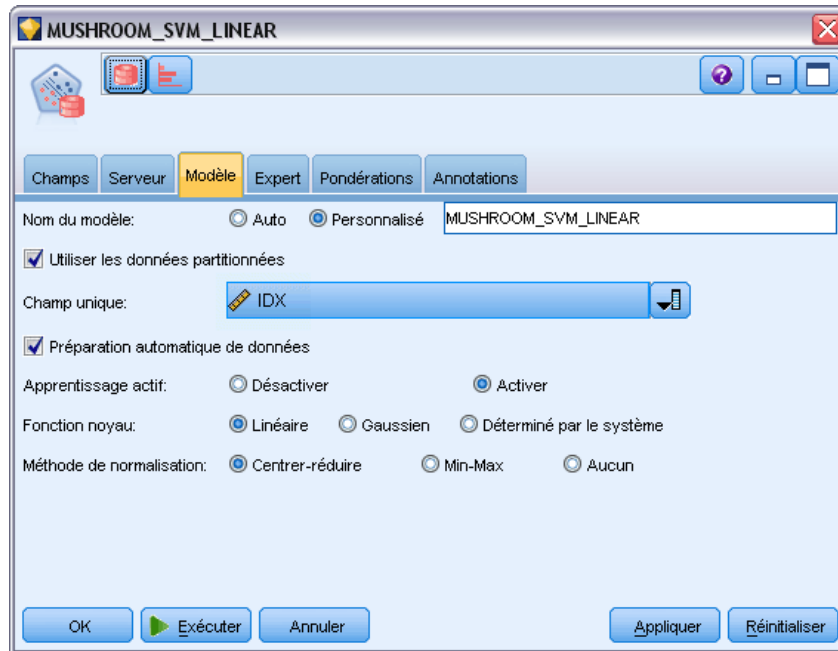
Support Vector Machine (SVM - Machine à vecteurs de prise en charge) est un algorithme de classification et de régression qui met en pratique la théorie de l'apprentissage automatique pour optimiser la précision des prévisions sans pour autant surajuster les données. SVM utilise une transformation non linéaire facultative des données d'apprentissage, suivie de la recherche des équations de régression dans les données transformées pour séparer les classes (pour les cibles catégorielles) ou ajuster la cible (pour les cibles continues). La mise en oeuvre de SVM par Oracle permet de créer des modèles via l'un des deux noyaux disponibles : le noyau linéaire ou le noyau

gaussien. Le noyau linéaire omet complètement la transformation non linéaire, si bien que le modèle obtenu est, pour l'essentiel, un modèle de régression.

Pour plus d'informations, reportez-vous aux manuels *Oracle Data Mining Application Developer's Guide* et *Oracle Data Mining Concepts*.

## Options des modèles Oracle SVM

Figure 4-9  
Options des modèles SVM



**Nom du modèle.** Vous pouvez générer le nom du modèle automatiquement sur la base du champ cible ou ID (ou du type de modèle si aucun de ces champs n'est spécifié) ou spécifier un nom personnalisé.

**Champ unique.** Indique le champ permettant d'identifier chaque observation de façon unique. Par exemple, il peut s'agir d'un champ d'ID, comme *ID client*. IBM® SPSS® Modeler impose une restriction selon laquelle ce champ-clé doit être numérique.

*Remarque :* Ce champ est en option pour tous les noeuds Oracle à l'exception des noeuds Oracle Adaptive Bayes, Oracle O-Cluster et Oracle Apriori.

**Préparation automatique de données.** (11g uniquement) Active (par défaut) ou désactive le mode de préparation automatique de données d'Oracle Data Mining. Si cette case est cochée, ODM effectue automatiquement les transformations de données nécessaires à l'algorithme. Pour des informations supplémentaires, consultez le manuel *Oracle Data Mining Concepts*.

**Apprentissage actif.** Permet de traiter de grands ensembles de création. Avec l'apprentissage actif, l'algorithme crée un modèle initial basé sur un petit échantillon avant de l'appliquer à l'ensemble de données d'apprentissage complet, puis met à jour l'échantillon et le modèle de manière

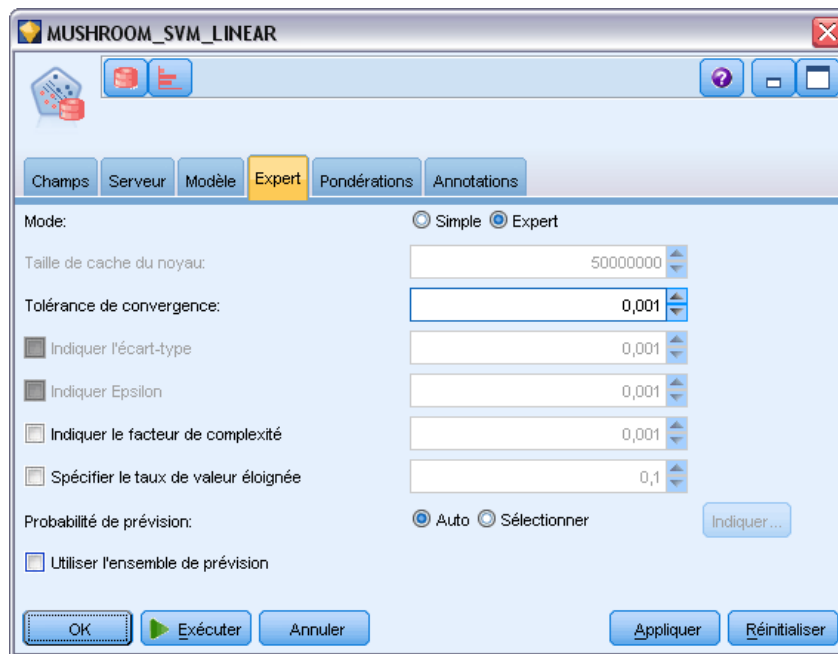
incrémentielle en fonction des résultats. Le cycle se répète jusqu'à ce que le modèle converge vers les données d'apprentissage ou que le nombre maximal de vecteurs de prise en charge soit atteint.

**Fonction noyau.** Sélectionnez Linéaire ou Gaussien, ou laissez le paramètre par défaut Système défini pour laisser le système choisir le noyau le plus adapté. Les noyaux gaussiens permettent d'étudier des relations plus complexes, mais les calculs effectués sont généralement plus longs. Vous pouvez commencer par le noyau linéaire et utiliser le noyau gaussien seulement si le noyau linéaire ne parvient pas à trouver un ajustement convenable. Ce cas de figure est plus probable avec un modèle de régression pour lequel le choix du noyau est plus important. En outre, notez que vous ne pouvez pas parcourir dans SPSS Modeler les modèles SVM créés par le biais du noyau gaussien. En revanche, vous pouvez le faire dans SPSS Modeler avec les modèles générés à l'aide du noyau linéaire de la même façon qu'avec les modèles de régression standard.

**Méthode de normalisation :** Indique la méthode de normalisation utilisée pour les champs cible et d'entrée continus. Vous pouvez sélectionner Centrer-réduire, Min-Max ou Aucun. Si la case Préparation automatique de données est cochée, Oracle effectue automatiquement la normalisation. Décochez cette case pour sélectionner manuellement la méthode de normalisation.

## Options expert Oracle SVM

Figure 4-10  
Options expert SVM



**Taille de cache du noyau.** Indique la taille (en octets) du cache à utiliser pour stocker les noyaux calculés lors de l'opération de création. En toute logique, un cache volumineux accélère généralement le processus de création. La valeur par défaut est 50 Mo.

**Tolérance de convergence.** Indique la valeur de tolérance autorisée avant la fin de la création du modèle. Cette valeur doit être comprise entre 0 et 1. La valeur par défaut est 0,001. Les valeurs supérieures ont tendance à entraîner une création rapide, mais génèrent des modèles moins précis.

**Indiquer l'écart-type.** Indique le paramètre d'écart-type utilisé par le noyau gaussien. Ce paramètre influe sur le compromis qu'implique la complexité du modèle et sa capacité à s'étendre à d'autres ensembles de données (surajustement et sous-ajustement des données). Les valeurs d'écart-type supérieures favorisent le sous-ajustement. Par défaut, ce paramètre est estimé à partir des données d'apprentissage.

**Indiquer Epsilon.** Pour les modèles de régression uniquement, indique la valeur de l'intervalle de l'erreur autorisée durant la création de modèles sur lesquels Epsilon n'a pas d'influence. En d'autres termes, il distingue les petites erreurs (non prises en compte) des erreurs importantes (prises en considération). Cette valeur doit être comprise entre 0 et 1. Par défaut, elle est estimée à partir des données d'apprentissage.

**Indiquer le facteur de complexité.** Indique le facteur de complexité, qui établit un compromis entre l'erreur d'un modèle (mesurée en fonction des données d'apprentissage) et la complexité de ce dernier, pour éviter tout surajustement ou sous-ajustement des données. Les valeurs supérieures sanctionnent davantage les erreurs et augmentent le risque de surajustement des données. Les valeurs inférieures, quant à elles, sanctionnent moins les erreurs et peuvent provoquer un sous-ajustement des données.

**Spécifier le taux de valeur éloignée.** Indique le taux de valeurs éloignées souhaité dans les données d'apprentissage. Valable uniquement pour les modèles SVM à classe unique. Ne peut pas être utilisé avec le paramètre **Indiquer le facteur de complexité**.

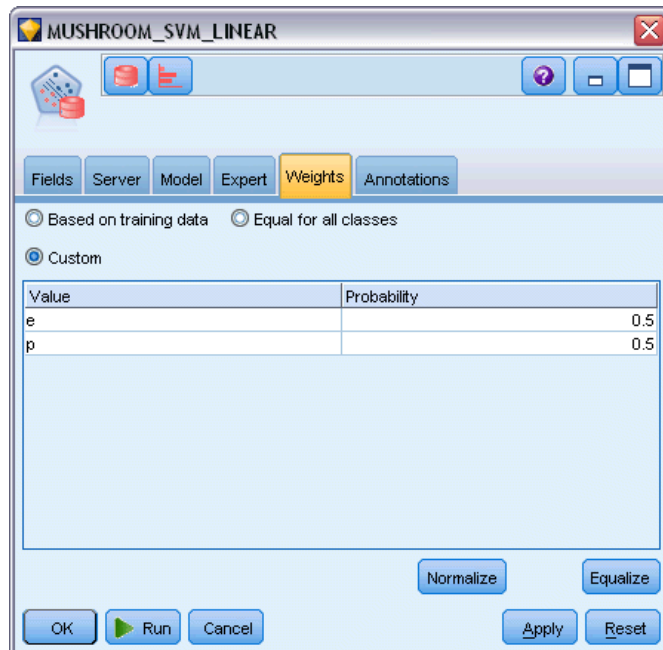
**Probabilité de prévision.** Permet au modèle d'inclure la probabilité d'une prévision correcte pour un résultat possible du champ cible. Pour activer cette fonction, sélectionnez Sélectionner, cliquez sur le bouton Spécifier, choisissez l'un des résultats possibles et cliquez sur Insérer.

**Utiliser l'ensemble de prévision.** Génère un tableau des résultats possibles du champ cible.



## Options de pondérations Oracle SVM

Figure 4-11  
Options de pondérations SVM



Dans un modèle de classification, l'utilisation des pondérations permet de définir l'importance relative des différentes valeurs cibles possibles. Par exemple, cela peut être utile si les points de données de vos données d'apprentissage ne sont pas distribués de manière réaliste entre les catégories. Les pondérations permettent d'orienter le modèle afin de compenser les catégories les moins représentées dans les données. L'augmentation de la pondération d'une valeur cible doit augmenter le pourcentage de prédictions correctes de cette catégorie.

Vous pouvez configurer les pondérations de trois façons différentes :

- **En fonction des données d'apprentissage.** Il s'agit de l'option par défaut. Les probabilités sont basées sur les effectifs relatifs des catégories dans les données d'apprentissage.
- **Identiques pour toutes les classes.** Pour toutes les catégories les pondérations sont définies par  $1/k$ , où  $k$  correspond au nombre de catégories cible.
- **Personnalisé.** Vous pouvez spécifier vos propres pondérations. Les valeurs de départ des pondérations sont définies comme étant identiques pour toutes les classes. Vous pouvez ensuite ajuster les pondérations de chaque catégorie selon les valeurs définies par l'utilisateur. Pour ajuster la pondération d'une catégorie spécifique, sélectionnez la cellule de pondération dans le tableau correspondant à la catégorie choisie, supprimez le contenu de la cellule et entrez la valeur souhaitée.

Le total des pondérations de toutes les catégories doit être égal à 1.0. Dans le cas contraire, un avertissement apparaît, avec une option d'effectuer un ajustement automatique des valeurs. Cette fonction d'ajustement automatique permet de préserver les proportions entre catégories tout en respectant la contrainte de pondération. Vous pouvez effectuer cet ajustement à tout moment en

cliquant sur le bouton Normaliser. Pour restaurer le tableau afin d'obtenir des valeurs égales dans toutes les catégories, cliquez sur le bouton Egaliser.

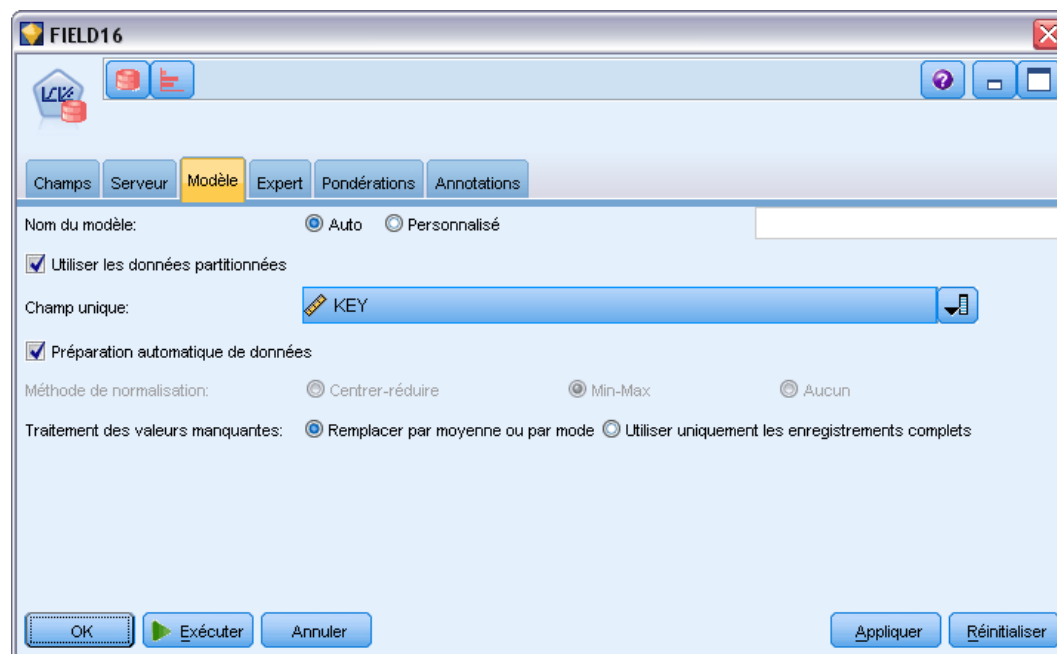
## Modèles linéaires généralisés (MLG) Oracle

(11g uniquement) Les modèles linéaires généralisés assouplissent les hypothèses restrictives effectuées par les modèles linéaires. Celles-ci comprennent par exemple les hypothèses que la variable cible a une distribution normale et que l'effet des variables indépendantes sur la variable cible est de nature linéaire. Un modèle linéaire généralisé est adapté aux hypothèses où la distribution de la cible peut être une distribution non normale, comme une loi multinomiale ou de Poisson. De même, un modèle linéaire généralisé est utile dans les cas où la relation ou le lien entre les variables indépendantes et la cible peut être non linéaire.

Pour plus d'informations, reportez-vous aux manuels *Oracle Data Mining Application Developer's Guide* et *Oracle Data Mining Concepts*.

### Options des modèles Oracle MLG

Figure 4-12  
options de modèle MLG



**Nom du modèle.** Vous pouvez générer le nom du modèle automatiquement sur la base du champ cible ou ID (ou du type de modèle si aucun de ces champs n'est spécifié) ou spécifier un nom personnalisé.

**Champ unique.** Indique le champ permettant d'identifier chaque observation de façon unique. Par exemple, il peut s'agir d'un champ d'ID, comme *ID client*. IBM® SPSS® Modeler impose une restriction selon laquelle ce champ-clé doit être numérique.

*Remarque* : Ce champ est en option pour tous les noeuds Oracle à l'exception des noeuds Oracle Adaptive Bayes, Oracle O-Cluster et Oracle Apriori.

**Préparation automatique de données.** (11g uniquement) Active (par défaut) ou désactive le mode de préparation automatique de données d'Oracle Data Mining. Si cette case est cochée, ODM effectue automatiquement les transformations de données nécessaires à l'algorithme. Pour des informations supplémentaires, consultez le manuel *Oracle Data Mining Concepts*.

**Méthode de normalisation** : Indique la méthode de normalisation utilisée pour les champs cible et d'entrée continus. Vous pouvez sélectionner Centrer-réduire, Min-Max ou Aucun. Si la case Préparation automatique de données est cochée, Oracle effectue automatiquement la normalisation. Décochez cette case pour sélectionner manuellement la méthode de normalisation.

**Gestion des valeurs manquantes.** Spécifie comment traiter les valeurs manquantes dans les données d'entrée :

- Remplacer par moyenne ou mode remplace les valeurs manquantes des attributs numériques par la valeur moyenne et remplace les valeurs manquantes des attributs catégoriels par le mode.
- N'utiliser que les enregistrements complets ignore les enregistrements avec des valeurs manquantes.

## Options expert Oracle MLG

Figure 4-13  
Options expert MLG

The screenshot shows the 'FIELD16' dialog box with the 'Expert' tab active. The 'Mode' section has 'Expert' selected. The 'Utiliser les pondérations de lignes' checkbox is checked. The 'Niveau de confiance du coefficient' is set to 0,95. The 'Catégorie de référence de la cible' is set to 'Personnalisé'. The 'Régression pseudo-orthogonale' is set to 'Désactiver'. The 'Probabilité de prévision' is set to 'Sélectionner'. Buttons for 'OK', 'Exécuter', 'Annuler', 'Appliquer', and 'Réinitialiser' are visible at the bottom.

**Utiliser les poids des lignes.** Cochez cette case pour activer la liste déroulante adjacente, dans laquelle vous pouvez sélectionner une colonne contenant un facteur de pondération pour les lignes.

**Enregistrer les diagnostics de ligne dans le tableau.** Cochez cette case pour activer le champ de texte adjacent, où vous pouvez saisir le nom d'un tableau qui contiendra les diagnostics des lignes.

**Coefficient de niveau de confiance.** Le degré de certitude, de 0.0 à 1.0, que la valeur prédite pour la cible se trouvera dans un intervalle de confiance calculé par le modèle. Les limites de confiance sont renvoyées avec les statistiques des coefficients.

**Catégorie de référence pour la cible.** Sélectionnez Personnalisé pour choisir une valeur pour le champ cible à utiliser comme catégorie de référence, ou laissez la valeur par défaut Auto .

**Régression pseudo-orthogonale.** La régression pseudo-orthogonale est une technique qui compense la situation dans laquelle le degré de corrélation des variables est trop élevé. Vous pouvez utiliser l'option Auto pour permettre à l'algorithme de contrôler l'utilisation de cette technique, ou vous pouvez la contrôler manuellement à l'aide des options Désactiver et Activer. Si vous choisissez d'activer manuellement la régression pseudo-orthogonale, vous pouvez remplacer la valeur par défaut du système du paramètre pseudo-orthogonal en entrant une valeur dans le champ adjacent.

**Produire un FIV pour la régression pseudo-orthogonale.** Cochez cette case si vous souhaitez produire des statistiques de Facteur d'Inflation de Variance (FIV) quand la méthode pseudo-orthogonale est utilisée pour la régression linéaire.

**Probabilité de prévision.** Permet au modèle d'inclure la probabilité d'une prévision correcte pour un résultat possible du champ cible. Pour activer cette fonction, sélectionnez Sélectionner, cliquez sur le bouton Spécifier, choisissez l'un des résultats possibles et cliquez sur Insérer.

**Utiliser l'ensemble de prévision.** Génère un tableau des résultats possibles du champ cible.

## Options de pondération Oracle MLG

Figure 4-14  
options de pondération MLG

The screenshot shows the 'FIELD16' dialog box with the 'Pondérations' tab selected. The 'Personnalisé' radio button is selected. The table below shows the following data:

Valeur	Probabilité
+	0,5
-	0,5

Buttons at the bottom include: OK, Exécuter, Annuler, Normaliser, Egaliser, Appliquer, Réinitialiser.

Dans un modèle de classification, l'utilisation des pondérations permet de définir l'importance relative des différentes valeurs cibles possibles. Par exemple, cela peut être utile si les points de données de vos données d'apprentissage ne sont pas distribués de manière réaliste entre les catégories. Les pondérations permettent d'orienter le modèle afin de compenser les catégories les moins représentées dans les données. L'augmentation de la pondération d'une valeur cible doit augmenter le pourcentage de prédictions correctes de cette catégorie.

Vous pouvez configurer les pondérations de trois façons différentes :

- **En fonction des données d'apprentissage.** Il s'agit de l'option par défaut. Les probabilités sont basées sur les effectifs relatifs des catégories dans les données d'apprentissage.
- **Identiques pour toutes les classes.** Pour toutes les catégories les pondérations sont définies par  $1/k$ , où  $k$  correspond au nombre de catégories cible.
- **Personnalisé.** Vous pouvez spécifier vos propres pondérations. Les valeurs de départ des pondérations sont définies comme étant identiques pour toutes les classes. Vous pouvez ensuite ajuster les pondérations de chaque catégorie selon les valeurs définies par l'utilisateur. Pour ajuster la pondération d'une catégorie spécifique, sélectionnez la cellule de pondération dans le tableau correspondant à la catégorie choisie, supprimez le contenu de la cellule et entrez la valeur souhaitée.

Le total des pondérations de toutes les catégories doit être égal à 1.0. Dans le cas contraire, un avertissement apparaît, avec une option d'effectuer un ajustement automatique des valeurs. Cette fonction d'ajustement automatique permet de préserver les proportions entre catégories tout en respectant la contrainte de pondération. Vous pouvez effectuer cet ajustement à tout moment en cliquant sur le bouton Normaliser. Pour restaurer le tableau afin d'obtenir des valeurs égales dans toutes les catégories, cliquez sur le bouton Egaliser.

## **Arbre décision Oracle**

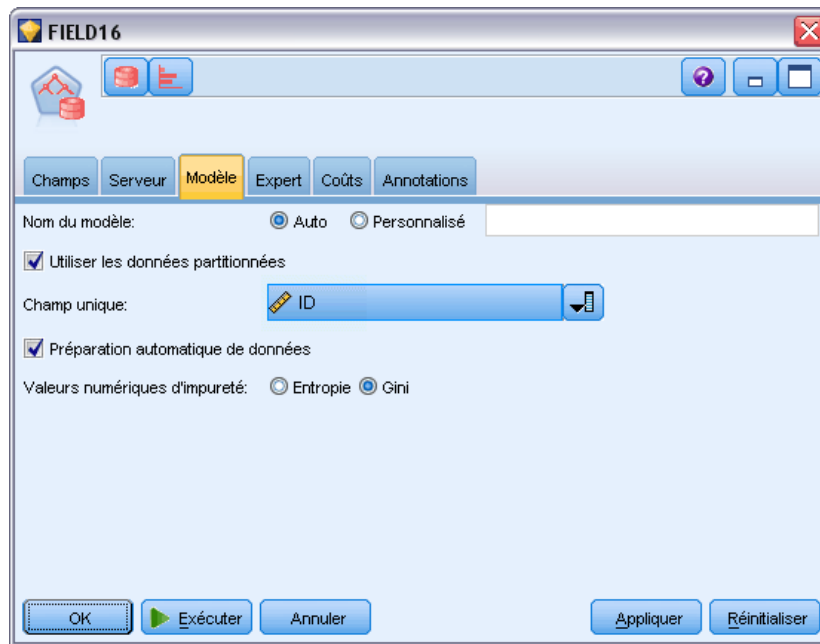
Oracle Data Mining offre une fonction d'arbre décision classique, qui s'appuie sur l'algorithme de l'arbre de classification supervisée et de régression CART (Classification and Regression Tree algorithm). Le modèle d'arbre décision ODM contient des informations complètes sur chaque noeud, y compris les noeuds Confiance, Prise en charge et Critère de division. La règle complète de chaque noeud peut être affichée. D'autre part, un attribut de substitution est fourni pour chacun des noeuds qui sera utilisé comme substitution lorsque le modèle sera appliqué à une observation dont les valeurs sont manquantes.

Les arbres décision sont populaires car ils sont applicables universellement, sont simples à comprendre et faciles à mettre en pratique. Les arbres décision examinent chaque attribut d'entrée potentiel en recherchant le meilleur "élément de fractionnement," à savoir, le point de césure d'un attribut ( $AGE > 55$ , par exemple) qui fractionne les enregistrements de données en aval en populations plus homogènes. Après chaque décision de fractionnement, ODM répète le processus en développant l'arbre complet et en créant les feuilles terminales qui représentent des populations similaires d'enregistrements, d'éléments ou de personnes. A partir du noeud d'arbre racine (par exemple, la population totale), les arbres décision fournissent des règles interprétables d'instructions IF A, then B. Ces règles d'arbre décision fournissent également la prise en charge et la confiance pour chaque noeud d'arbre.

Les réseaux Adaptive Bayes peuvent fournir des règles simples et courtes pouvant servir à expliquer chaque prévision, mais les arbres décision, quant à eux, fournissent des règles complètes Oracle Data Mining pour chaque décision de division. Les arbres décision servent également à développer des profils détaillés des meilleurs clients, des patients en bonne santé, des facteurs liés à la fraude, etc.

## Options du modèle Arbre décision

Figure 4-15  
Options du modèle Arbre décision



**Nom du modèle.** Vous pouvez générer le nom du modèle automatiquement sur la base du champ cible ou ID (ou du type de modèle si aucun de ces champs n'est spécifié) ou spécifier un nom personnalisé.

**Champ unique.** Indique le champ permettant d'identifier chaque observation de façon unique. Par exemple, il peut s'agir d'un champ d'ID, comme *ID client*. IBM® SPSS® Modeler impose une restriction selon laquelle ce champ-clé doit être numérique.

*Remarque :* Ce champ est en option pour tous les noeuds Oracle à l'exception des noeuds Oracle Adaptive Bayes, Oracle O-Cluster et Oracle Apriori.

**Préparation automatique de données.** (11g uniquement) Active (par défaut) ou désactive le mode de préparation automatique de données d'Oracle Data Mining. Si cette case est cochée, ODM effectue automatiquement les transformations de données nécessaires à l'algorithme. Pour des informations supplémentaires, consultez le manuel *Oracle Data Mining Concepts*.

**Valeurs numériques d'impureté.** Indique les valeurs numériques utilisées pour rechercher la meilleure question test permettant de fractionner les données dans chaque noeud. L'élément de fractionnement et la valeur de fraction les plus appropriés sont ceux qui induisent la plus grande

augmentation de l'homogénéité de la valeur cible pour les entités du noeud. L'homogénéité est mesurée par rapport à une valeur numérique. Les valeurs numériques prises en charge sont **gini** et **entropy**.

## Arbre de décision - Options expert

Figure 4-16  
Arbre de décision - Options expert



**Profondeur maximale.** Définit la profondeur maximale du modèle d'arbre à créer.

**Pourcentage minimal d'enregistrements dans un noeud.** Définit le pourcentage du nombre minimal d'enregistrements par noeud.

**Pourcentage minimal d'enregistrements pour une division.** Définit le nombre minimal d'enregistrements dans un noeud parent, exprimé en pourcentage du nombre total d'enregistrements utilisés pour l'apprentissage du modèle. Si le nombre d'enregistrements est inférieur à ce pourcentage, aucune division n'est appliquée.

**Enregistrements minimaux dans un noeud.** Définit le nombre minimal d'enregistrements renvoyés.

**Enregistrements minimaux pour une division.** Définit le nombre minimal d'enregistrements dans un noeud parent, exprimé en valeur. Si le nombre d'enregistrements est inférieur à cette valeur, aucune division n'est appliquée.

**Identificateur de règle.** S'il est activé, l'identificateur insère une chaîne dans le modèle qui permet d'identifier le noeud dans l'arbre auquel une division spécifique a lieu.

**Probabilité de prévision.** Permet au modèle d'inclure la probabilité d'une prévision correcte pour un résultat possible du champ cible. Pour activer cette fonction, sélectionnez Sélectionner, cliquez sur le bouton Spécifier, choisissez l'un des résultats possibles et cliquez sur Insérer.

**Utiliser l'ensemble de prévision.** Génère un tableau des résultats possibles du champ cible.

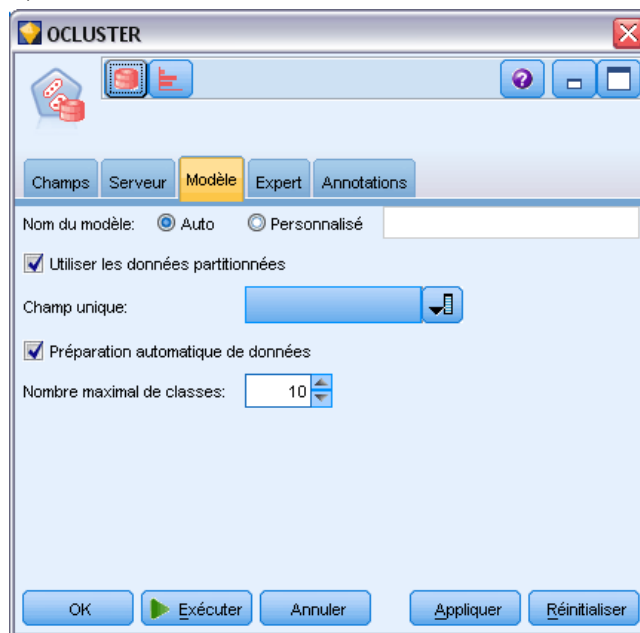
## O-Cluster Oracle

L'algorithme O-Cluster Oracle identifie les regroupements naturels au sein d'une population de données. La classification non supervisée par partition orthogonale (O-Cluster) est un algorithme de classification exclusif d'Oracle qui crée un modèle de classification hiérarchique de type grille. En d'autres termes, il crée des partitions parallèles à l'axe (orthogonales) dans l'espace des attributs d'entrée. L'algorithme fonctionne de manière réursive. La structure hiérarchique résultante représente une grille irrégulière qui génère un pavage en classes de l'espace des attributs.

L'algorithme O-Cluster gère les attributs numériques et catégoriels, et Oracle Data Mining (ODM) sélectionne automatiquement les meilleures définitions de classe. ODM fournit des informations détaillées sur les classes, leurs règles et leurs valeurs centroïdes, et permet d'évaluer une population sur la base de son appartenance à une classe.

### Options du modèle O-Cluster

Figure 4-17  
Options du modèle O-Cluster



**Nom du modèle.** Vous pouvez générer le nom du modèle automatiquement sur la base du champ cible ou ID (ou du type de modèle si aucun de ces champs n'est spécifié) ou spécifier un nom personnalisé.

**Champ unique.** Indique le champ permettant d'identifier chaque observation de façon unique. Par exemple, il peut s'agir d'un champ d'ID, comme *ID client*. IBM® SPSS® Modeler impose une restriction selon laquelle ce champ-clé doit être numérique.



*Remarque* : Ce champ est en option pour tous les noeuds Oracle à l'exception des noeuds Oracle Adaptive Bayes, Oracle O-Cluster et Oracle Apriori.

**Préparation automatique de données.** (11g uniquement) Active (par défaut) ou désactive le mode de préparation automatique de données d'Oracle Data Mining. Si cette case est cochée, ODM effectue automatiquement les transformations de données nécessaires à l'algorithme. Pour des informations supplémentaires, consultez le manuel *Oracle Data Mining Concepts*.

**Nombre maximal de règles.** Définit le nombre maximum de classes générées.

## Options expert O-Cluster

Figure 4-18  
Options expert O-Cluster



**Mémoire tampon maximale.** Définit la taille de tampon maximale.

**Sensibilité.** Définit une fraction qui indique la densité maximale requise pour séparer une nouvelle classe. La fraction est liée à la densité uniforme globale.

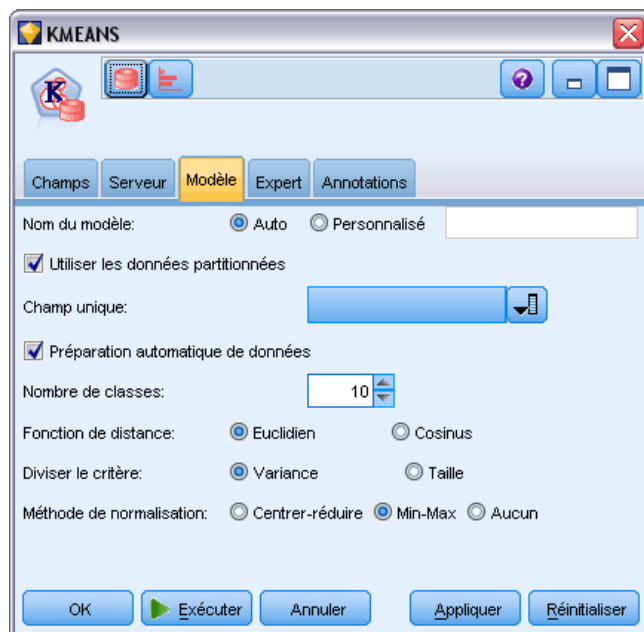
## Algorithme Oracle k-Means

L'algorithme Oracle k-Means identifie les regroupements naturels au sein d'une population de données. L'algorithme k-Means est un algorithme de classification non supervisée basé sur la distance qui partitionne les données en un nombre prédéterminé de classes (à condition que le nombre d'observations distinctes soit suffisant). Les algorithmes basés sur la distance dépendent d'une mesure (fonction) de la distance pour mesurer la similarité entre des points de données. Les points de données sont affectés à la classe la plus proche, en fonction de la mesure de distance utilisée. ODM fournit une version améliorée de k-Means.

L'algorithme k-Means prend en charge les classes hiérarchiques, gère les attributs numériques et catégoriels et répartit la population en un nombre de classes défini par l'utilisateur. ODM fournit des informations détaillées sur les classes, leurs règles et leurs valeurs centroïdes, et permet d'évaluer une population sur la base de son appartenance à une classe.

## Options du modèle k-Means

Figure 4-19  
Options du modèle k-Means



**Nom du modèle.** Vous pouvez générer le nom du modèle automatiquement sur la base du champ cible ou ID (ou du type de modèle si aucun de ces champs n'est spécifié) ou spécifier un nom personnalisé.

**Champ unique.** Indique le champ permettant d'identifier chaque observation de façon unique. Par exemple, il peut s'agir d'un champ d'ID, comme *ID client*. IBM® SPSS® Modeler impose une restriction selon laquelle ce champ-clé doit être numérique.

*Remarque :* Ce champ est en option pour tous les noeuds Oracle à l'exception des noeuds Oracle Adaptive Bayes, Oracle O-Cluster et Oracle Apriori.

**Préparation automatique de données.** (11g uniquement) Active (par défaut) ou désactive le mode de préparation automatique de données d'Oracle Data Mining. Si cette case est cochée, ODM effectue automatiquement les transformations de données nécessaires à l'algorithme. Pour des informations supplémentaires, consultez le manuel *Oracle Data Mining Concepts*.

**Nombre de classes.** Définit le nombre de classes générées.

**Fonction de distance.** Indique quelle fonction de distance est utilisée pour la classification non supervisée k-Means.

**Diviser le critère.** Indique quel critère de division est utilisé pour la classification non supervisée k-Means.

**Méthode de normalisation.** Indique la méthode de normalisation utilisée pour les champs cible et d'entrée continus. Vous pouvez sélectionner Centrer-réduire, Min-Max ou Aucun.

### Options expert de k-Means

Figure 4-20  
Options expert de k-Means



**Itérations :** Définit le nombre d'itérations pour l'algorithme k-Means.

**Tolérance de convergence.** Définit la tolérance de convergence pour l'algorithme k-Means.

**Nombre d'intervalles.** Définit le nombre d'intervalles dans l'histogramme d'attributs produit par k-Means. Les limites d'intervalle de chaque attribut sont calculées de manière globale, sur l'intégralité de l'ensemble de données d'apprentissage. La méthode consiste à créer des intervalles de largeur uniforme (equi-width). Tous les attributs ont le même nombre d'intervalles, à l'exception des attributs à valeur unique, qui n'en ont qu'un seul.

**Déploiement de bloc.** Définit le facteur de croissance de la mémoire allouée aux données de classe.

**Prise en charge d'attribut de pourcentage minimal.** Définit la fraction des valeurs d'attribut qui doivent être non nulles pour que l'attribut soit inclus dans la description de règle de la classe. L'attribution d'une valeur trop élevée au paramètre, lorsque des valeurs manquent au sein des données, peut générer des règles très courtes, voire vides.

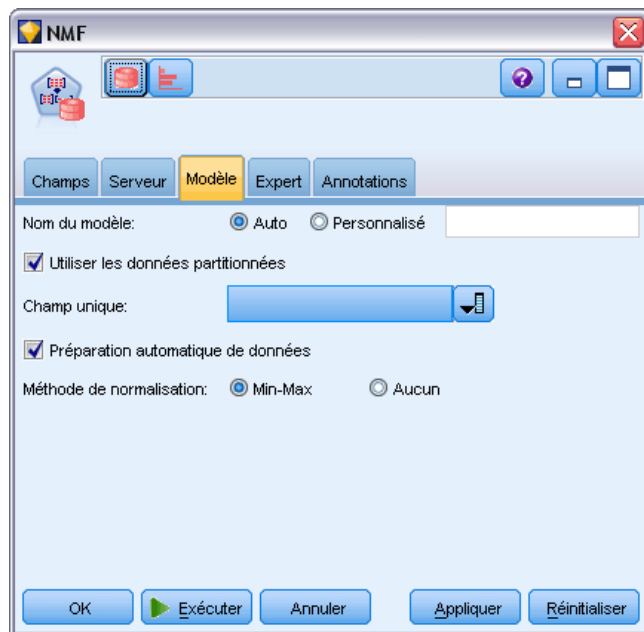
## Oracle Nonnegative Matrix Factorization (NMF)

La factorisation en matrices non négatives (Nonnegative Matrix Factorization, NMF) est utile pour réduire un ensemble de données volumineux en attributs représentatifs. De concept semblable à l'analyse en composantes principales (ACP), la NMF est cependant capable de gérer des quantités d'attributs plus importantes, dans un modèle de représentation additif. La NMF est un puissant algorithme de Data mining d'avant-garde, qui peut être exécuté dans divers cas d'utilisation.

La NMF permet de réduire les données volumineuses, par exemple des données texte, en représentations plus petites et plus sporadiques qui limitent la dimensionnalité des données (les mêmes informations peuvent être préservées avec bien moins de variables). Les résultats des modèles NMF peuvent être analysés à l'aide de techniques d'apprentissage supervisées, telles que les SVM, ou de techniques non supervisées, telles que les techniques de classification. Oracle Data Mining utilise les algorithmes NMF et SVM pour explorer les données texte non structurées.

### Options du modèle NMF

Figure 4-21  
Options du modèle NMF



**Nom du modèle.** Vous pouvez générer le nom du modèle automatiquement sur la base du champ cible ou ID (ou du type de modèle si aucun de ces champs n'est spécifié) ou spécifier un nom personnalisé.

**Champ unique.** Indique le champ permettant d'identifier chaque observation de façon unique. Par exemple, il peut s'agir d'un champ d'ID, comme *ID client*. IBM® SPSS® Modeler impose une restriction selon laquelle ce champ-clé doit être numérique.

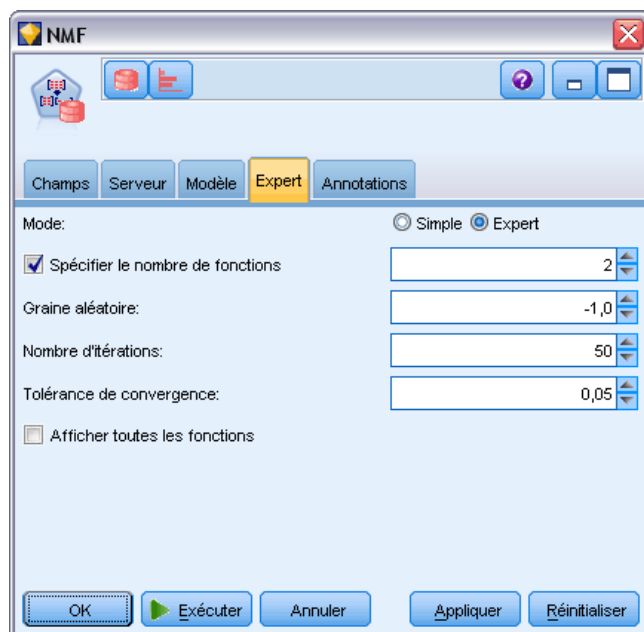
*Remarque :* Ce champ est en option pour tous les noeuds Oracle à l'exception des noeuds Oracle Adaptive Bayes, Oracle O-Cluster et Oracle Apriori.

**Préparation automatique de données.** (11g uniquement) Active (par défaut) ou désactive le mode de préparation automatique de données d'Oracle Data Mining. Si cette case est cochée, ODM effectue automatiquement les transformations de données nécessaires à l'algorithme. Pour des informations supplémentaires, consultez le manuel *Oracle Data Mining Concepts*.

**Méthode de normalisation :** Indique la méthode de normalisation utilisée pour les champs cible et d'entrée continus. Vous pouvez sélectionner Centrer-réduire, Min-Max ou Aucun. Si la case Préparation automatique de données est cochée, Oracle effectue automatiquement la normalisation. Décochez cette case pour sélectionner manuellement la méthode de normalisation.

## Options expert NMF

Figure 4-22  
Options expert NMF



**Spécifier le nombre de fonctions.** Définit le nombre de fonctions à extraire.

**Graine aléatoire.** Définit la graine aléatoire de l'algorithme NMF.

**Nombre d'itérations.** Définit le nombre d'itérations pour l'algorithme NMF.

**Tolérance de convergence.** Définit la tolérance de convergence de l'algorithme NMF.

**Afficher toutes les fonctions.** Affiche l'ID et la confiance de toutes les fonctions, et non pas uniquement les valeurs de la meilleure fonction.

## **Apriori Oracle**

L'algorithme Apriori recherche des règles d'association dans les données. Par exemple, « si un client achète un rasoir et de l'après-rasage, alors la probabilité pour que ce client achète de la crème à raser est de 80 % ». Le problème d'exploration des associations peut être décomposé en deux sous-problèmes :

- Trouver toutes les combinaisons d'éléments, appelées jeux d'éléments fréquents, dont la prise en charge est supérieure à la prise en charge minimale.
- Utiliser les jeux d'éléments fréquents pour générer les règles souhaitées. Le concept est le suivant : supposons qu'ABC et BC sont fréquents. La règle « A implique BC » est alors vérifiée si le rapport  $\text{support}(ABC) / \text{support}(BC)$  est au moins aussi important que la confiance minimale. Notez que la règle aura la prise en charge minimale car ABCD est fréquent. L'association ODM ne prend en charge que les règles à conséquence unique (ABC implique D).

Le nombre de jeux d'éléments fréquents dépend des paramètres de prise en charge minimale. Le nombre de règles générées dépend du nombre de jeux d'éléments fréquents et du paramètre de confiance. Si le paramètre de confiance est trop élevé, le modèle d'association risque de contenir des jeux d'éléments fréquents mais pas de règles.

ODM utilise une mise en oeuvre SQL de l'algorithme Apriori. Les étapes de génération des candidats et de comptage des prises en charge sont mises en oeuvre à l'aide de requêtes SQL. Les structures de données spécialisées en mémoire ne sont pas utilisées. Pour une exécution plus efficace sur le serveur de base de données, les requêtes SQL sont affinées à l'aide de divers conseils.

### **Options des champs Apriori**

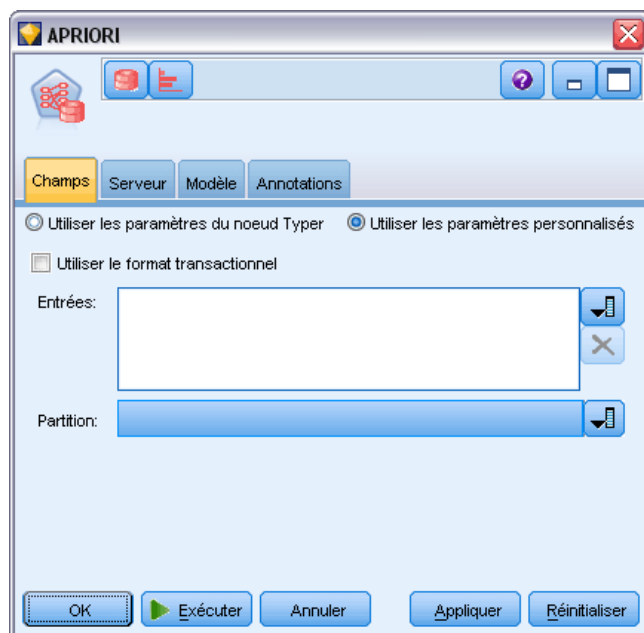
Tous les noeuds de modélisation comportent un onglet Champs, vous permettant de spécifier les champs à utiliser lors de la construction du modèle.

Avant de pouvoir créer un modèle Apriori, il est nécessaire de définir quels champs sont intéressants à utiliser pour la modélisation d'association.

**Utiliser les paramètres du noeud Typer.** Cette option indique au noeud d'utiliser les informations du champ à partir d'un noeud Typer en amont. Il s'agit de la valeur par défaut.

**Utiliser les paramètres personnalisés.** Cette option indique au noeud d'utiliser les informations du champ spécifiées ici au lieu des informations données dans un noeud Typer en amont. Après avoir sélectionné cette option, définissez les champs restants dans la boîte de dialogue selon si vous utilisez ou non le format transactionnel.

Figure 4-23  
Paramètres par défaut des champs personnalisés



Si vous n'utilisez *pas* le format transactionnel, choisissez :

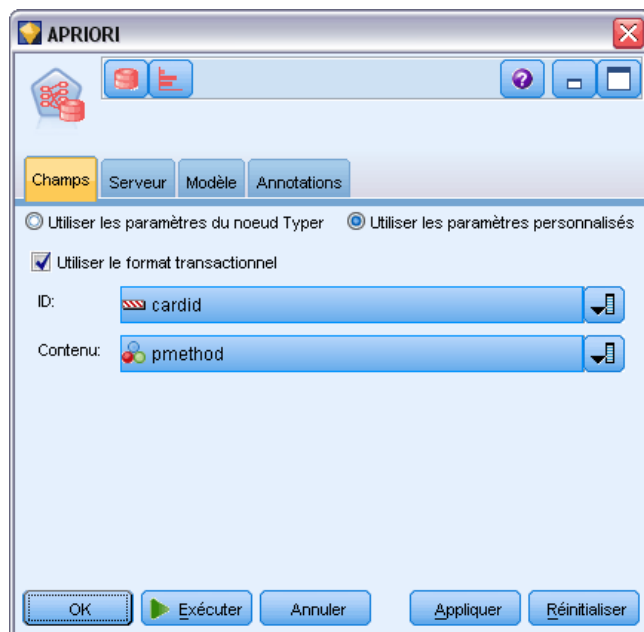
- **Entrées.** Sélectionnez le ou les champs d'entrée. Cela revient à définir le rôle du champ sur la valeur *Entrée* dans un nœud Typier.
- **Partition.** Ce champ permet d'indiquer un champ utilisé pour partitionner les données en échantillons distincts pour les étapes d'apprentissage, de test et de validation de la création d'un modèle. [Pour plus d'informations, reportez-vous à la section Nœud Partitionner dans le chapitre 4 dans \*Noeuds source, exécution et de sortie de IBM SPSS Modeler 15\*.](#)

Si vous *utilisez* le format transactionnel, choisissez :

**Utiliser le format transactionnel.** Utilisez cette option si vous souhaitez transformer les données d'une ligne par élément à une ligne par observation.

Sélectionner cette option modifie les commandes de champ dans la partie inférieure de cette boîte de dialogue :

Figure 4-24  
les paramètres de champ du format transactionnel



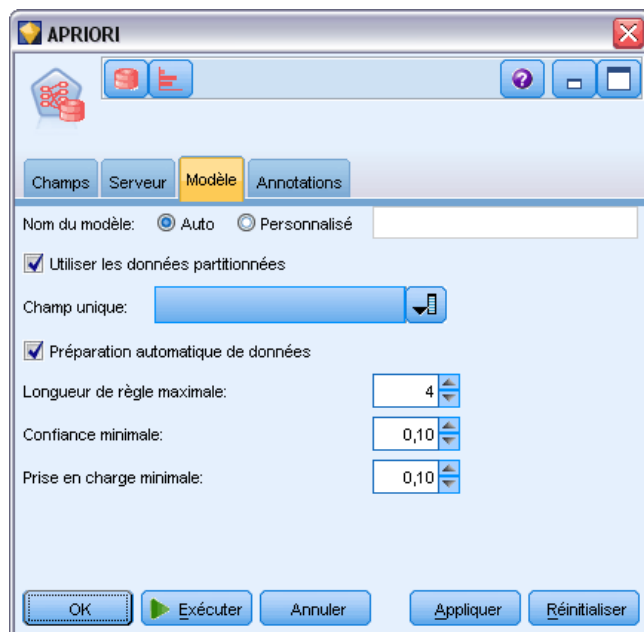
Pour le format transactionnel, spécifiez les éléments suivants :

- **ID.** Sélectionnez un champ ID dans la liste. Vous pouvez utiliser un champ numérique ou symbolique en tant que champ ID. Chaque valeur unique de ce champ doit indiquer une unité d'analyse spécifique. Par exemple, dans une analyse de paniers du marché, chaque ID peut représenter un client. Dans une analyse de l'utilisation du Web, chaque ID peut représenter un ordinateur (adresse IP) ou un utilisateur (ID de connexion).
- **Contenu.** Spécifiez le champ de contenu du modèle. Ce champ contient l'élément d'intérêt concernant la modélisation des associations.
- **Partition.** Ce champ permet d'indiquer un champ utilisé pour partitionner les données en échantillons distincts pour les étapes d'apprentissage, de test et de validation de la création d'un modèle. L'utilisation d'un échantillon pour la création du modèle et d'un échantillon distinct pour le tester vous permet d'avoir une bonne indication de la manière dont le modèle peut se généraliser à des ensembles de données plus importants, similaires aux données actuelles. Si plusieurs champs de partition sont définis via des noeuds Typier ou Partitionner, vous devez en sélectionner un seul dans l'onglet Champs de chaque noeud de modélisation ayant recours au partitionnement. (Dans le cas d'une seule partition, cette partition est automatiquement utilisée lorsque la fonction de partition est activée.) [Pour plus d'informations, reportez-vous à la section Noeud Partitionner dans le chapitre 4 dans Noeuds source, exécution et de sortie de IBM SPSS Modeler 15.](#) Notez également que, pour appliquer la partition sélectionnée à l'analyse, vous devez activer l'option de partitionnement dans l'onglet Options de modèle du noeud. (Désélectionnez cette option pour pouvoir désactiver la partition sans modifier les paramètres du champ.)



## Options du modèle Apriori

Figure 4-25  
Options du modèle Apriori



**Nom du modèle.** Vous pouvez générer le nom du modèle automatiquement sur la base du champ cible ou ID (ou du type de modèle si aucun de ces champs n'est spécifié) ou spécifier un nom personnalisé.

**Champ unique.** Indique le champ permettant d'identifier chaque observation de façon unique. Par exemple, il peut s'agir d'un champ d'ID, comme *ID client*. IBM® SPSS® Modeler impose une restriction selon laquelle ce champ-clé doit être numérique.

*Remarque :* Ce champ est en option pour tous les noeuds Oracle à l'exception des noeuds Oracle Adaptive Bayes, Oracle O-Cluster et Oracle Apriori.

**Préparation automatique de données.** (11g uniquement) Active (par défaut) ou désactive le mode de préparation automatique de données d'Oracle Data Mining. Si cette case est cochée, ODM effectue automatiquement les transformations de données nécessaires à l'algorithme. Pour des informations supplémentaires, consultez le manuel *Oracle Data Mining Concepts*.

**Longueur de règle maximale.** Définit le nombre maximum de pré-conditions pour toute règle, un entier entre 2 et 20. C'est une manière de limiter la complexité des règles. Si les règles sont trop complexes ou trop spécifiques, ou si votre ensemble de règles est trop long à apprendre, essayez de diminuer ce paramètre.

**Confiance minimale.** Définit le niveau de confiance minimum, une valeur entre 0 et 1. Les règles ayant une confiance inférieure au critère spécifié sont ignorées.

**Prise en charge minimale.** Définit le seuil de prise en charge minimum, une valeur entre 0 et 1. Apriori recherche les tendances ayant un effectif supérieur au seuil de prise en charge minimum.

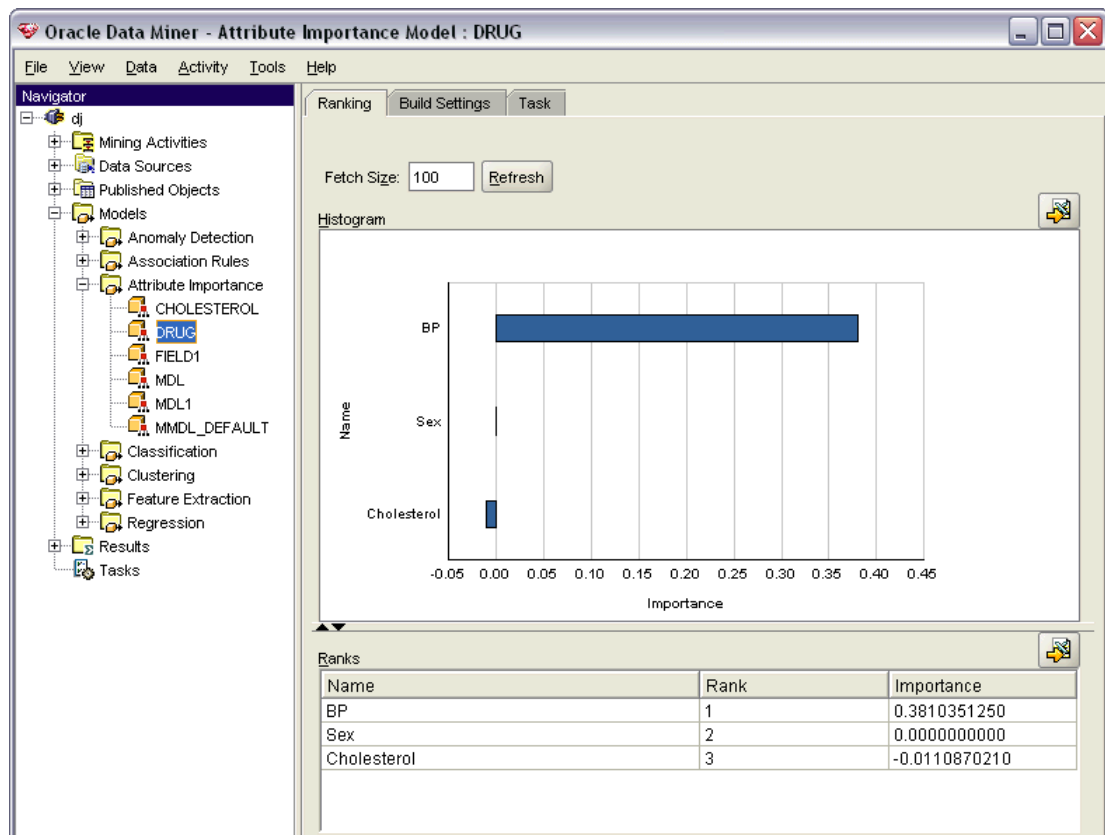
## Description de longueur minimale d'Oracle (MDL)

L'algorithme de description de longueur minimale d'Oracle (MDL) aide à identifier les attributs qui ont l'influence la plus importante sur un attribut cible. Très souvent, le fait de savoir quels attributs ont le plus d'influence vous aide à mieux comprendre et à mieux gérer le problème, ainsi qu'à simplifier les activités de modélisation. Par ailleurs, ces attributs peuvent indiquer les types de données que vous pouvez ajouter pour compléter vos modèles. La MDL permet, par exemple, de rechercher les attributs de processus les plus pertinents pour prédire la qualité d'une pièce fabriquée, les facteurs associés à l'attrition ou les gènes les plus susceptibles d'être pris en compte pour le traitement d'une maladie donnée.

La MDL d'Oracle ignore les champs d'entrée qu'elle considère comme inutiles pour prédire la cible. Avec les champs d'entrée restants, elle construit ensuite un nugget de modèle brut qui est associé à un modèle Oracle visible dans Oracle Data Miner. Si vous naviguez jusqu'au modèle dans Oracle Data Miner, un graphique apparaît avec les champs d'entrée restants, classés dans l'ordre de leur importance pour prédire la cible.

Figure 4-26

Utilisation d'un graphique MDL d'Oracle présentant l'importance relative des champs d'entrée dans la prédiction d'une cible



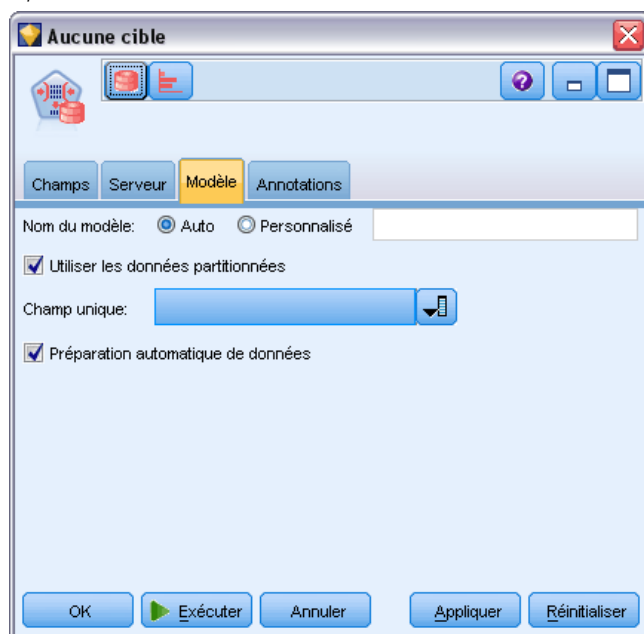
Un classement négatif est indicateur de parasites. Les champs d'entrée classés zéro ou moins ne contribuent pas aux prédictions et devraient probablement être supprimés des données.

**Pour afficher le graphique**

- ▶ Faites un clic droit sur le nugget du modèle brut dans la palette Modèles et choisissez Parcourir.
- ▶ Dans la fenêtre du modèle, cliquez sur le bouton pour lancer Oracle Data Miner.
- ▶ Connectez-vous à Oracle Data Miner. [Pour plus d'informations, reportez-vous à la section Oracle Data Miner sur p. 97.](#)
- ▶ Dans le panneau du navigateur d'Oracle Data Miner, développez Modèles, puis Importance de l'attribut.
- ▶ Sélectionnez le modèle Oracle approprié (il porte le même nom que le champ cible spécifié dans IBM® SPSS® Modeler). Si vous n'êtes pas certain du modèle, sélectionnez le dossier Importance de l'attribut et recherchez un modèle par date de création.

**Options du modèle MDL**

Figure 4-27  
Options du modèle MDL



**Nom du modèle.** Vous pouvez générer le nom du modèle automatiquement sur la base du champ cible ou ID (ou du type de modèle si aucun de ces champs n'est spécifié) ou spécifier un nom personnalisé.

**Champ unique.** Indique le champ permettant d'identifier chaque observation de façon unique. Par exemple, il peut s'agir d'un champ d'ID, comme *ID client*. IBM® SPSS® Modeler impose une restriction selon laquelle ce champ-clé doit être numérique.

*Remarque :* Ce champ est en option pour tous les noeuds Oracle à l'exception des noeuds Oracle Adaptive Bayes, Oracle O-Cluster et Oracle Apriori.

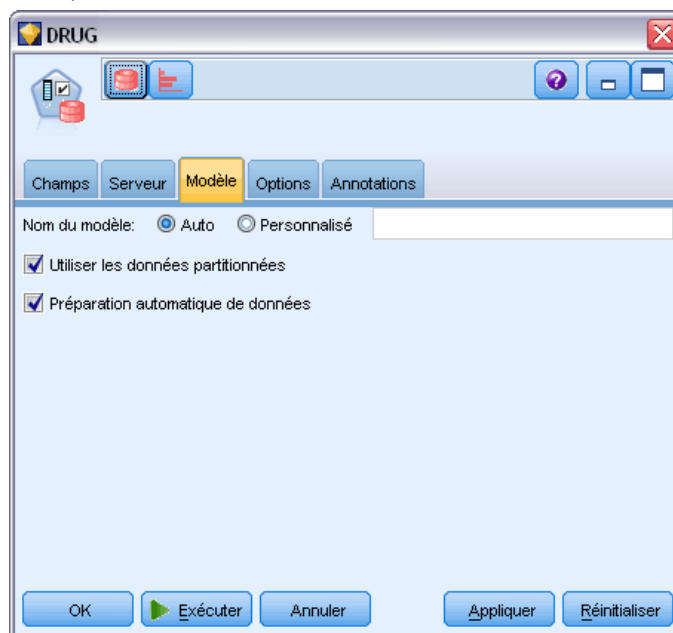
**Préparation automatique de données.** (11g uniquement) Active (par défaut) ou désactive le mode de préparation automatique de données d'Oracle Data Mining. Si cette case est cochée, ODM effectue automatiquement les transformations de données nécessaires à l'algorithme. Pour des informations supplémentaires, consultez le manuel *Oracle Data Mining Concepts*.

## Importance de l'attribut d'Oracle (IA)

L'objectif de l'importance de l'attribut est de découvrir quels attributs dans l'ensemble de données sont associés au résultat, et le degré d'influence qu'ils ont sur les résultats finaux. Le noeud Importance de l'attribut d'Oracle analyse les données, recherche les motifs et prédit les résultats avec un niveau de confiance associé.

### IA - Options du modèle

Figure 4-28  
IA - Options du modèle



**Nom du modèle.** Vous pouvez générer le nom du modèle automatiquement sur la base du champ cible ou ID (ou du type de modèle si aucun de ces champs n'est spécifié) ou spécifier un nom personnalisé.

**Utiliser les données partitionnées.** Si un champ de partition est défini, cette option assure que seules les données provenant de la partition d'apprentissage sont utilisées pour générer le modèle. [Pour plus d'informations, reportez-vous à la section Noeud Partitionner dans le chapitre 4 dans \*Noeuds source, exécution et de sortie de IBM SPSS Modeler 15\*.](#)

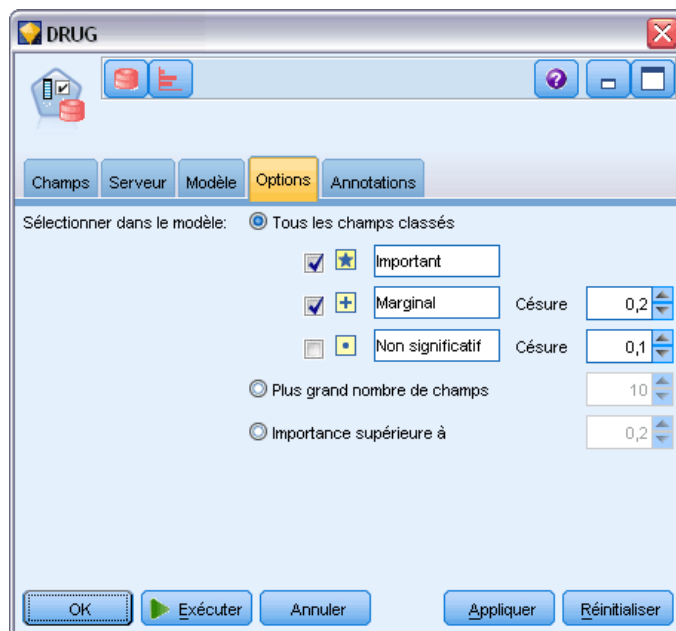
**Préparation automatique de données.** (11g uniquement) Active (par défaut) ou désactive le mode de préparation automatique de données d'Oracle Data Mining. Si cette case est cochée, ODM effectue automatiquement les transformations de données nécessaires à l'algorithme. Pour des informations supplémentaires, consultez le manuel *Oracle Data Mining Concepts*.

## IA - Options de sélection

L'onglet Options vous permet d'indiquer les paramètres par défaut pour la sélection et l'exclusion des champs d'entrée du nugget de modèle. Vous pouvez ensuite ajouter le modèle à un flux afin de sélectionner un sous-ensemble de champs, en vue de futures opérations de création de modèles. Il est également possible d'ignorer ces paramètres en sélectionnant ou désélectionnant d'autres champs dans le navigateur du modèle généré. Toutefois, les paramètres par défaut permettent d'appliquer, sans autres modifications, le nugget de modèle ; cela peut s'avérer particulièrement utile pour la génération de scripts.

Figure 4-29

IA - Options de sélection



Les options suivantes sont disponibles :

**Tous les champs classés.** Sélectionne les champs en fonction de leur classement et les indique comme étant *Important*, *Marginal* ou *Non significatif*. Vous pouvez éditer l'étiquette de chaque classement, ainsi que les valeurs de césure utilisées pour affecter les enregistrements à un rang ou un autre.

**Plus grand nombre de champs.** Sélectionne les  $n$  premiers champs en termes d'importance.

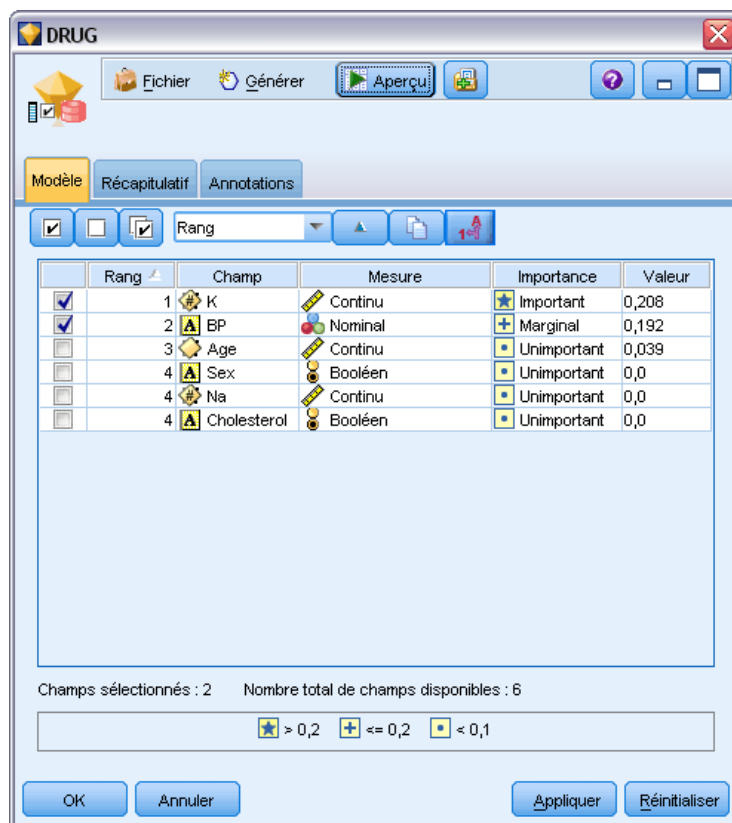
**Importance supérieure à.** Sélectionne tous les champs présentant une importance supérieure à la valeur indiquée.

Le champ cible est toujours conservé quelle que soit la sélection.

## Nugget du modèle IA - Onglet Modèle

L'onglet Modèle correspondant à un nugget de modèle IA d'Oracle affiche le classement et l'importance de toutes les entrées et vous permet également de sélectionner les champs de filtrage en cochant les cases situées dans la colonne de gauche. Lorsque vous exécutez le flux, seuls les champs sélectionnés sont conservés en plus des prédictions de cible. Les autres champs d'entrée sont ignorés. Les sélections par défaut dépendent des options indiquées dans le noeud de modélisation, mais vous pouvez sélectionner ou désélectionner d'autres champs, comme souhaité.

Figure 4-30  
Nugget du modèle IA



- Pour trier la liste par rang, nom de champ, importance ou toute autre colonne affichée, cliquez sur l'en-tête de colonne souhaité. Vous pouvez également sélectionner l'élément souhaité dans la liste à côté du bouton Trier par, et utiliser les flèches vers le haut ou vers le bas pour changer le sens du tri.
- Dans la barre d'outils, vous pouvez sélectionner ou désélectionner tous les champs, et accéder à la boîte de dialogue Sélectionner les champs pour sélectionner les champs par rang ou importance. Vous pouvez également appuyer sur les touches Maj ou Ctrl tout en cliquant sur les champs pour développer la sélection. [Pour plus d'informations, reportez-vous à la](#)

section *Sélection des champs en fonction de leur importance* dans le chapitre 4 dans *Noeuds de modélisation de IBM SPSS Modeler 15*.

- Les valeurs de seuil permettant de classer les entrées comme étant importantes, marginales ou non significatives sont affichées dans la légende apparaissant sous le tableau. Ces valeurs sont définies dans le noeud de modélisation.

## ***Gestion des modèles Oracle***

Les modèles Oracle sont ajoutés à la palette Modèles à l'instar des autres modèles IBM® SPSS® Modeler. En outre, ils peuvent également être utilisés quasiment de la même manière. Toutefois, on distingue quelques différences importantes, dans la mesure où chaque modèle Oracle créé dans SPSS Modeler fait en réalité référence à un modèle stocké sur un serveur de base de données.

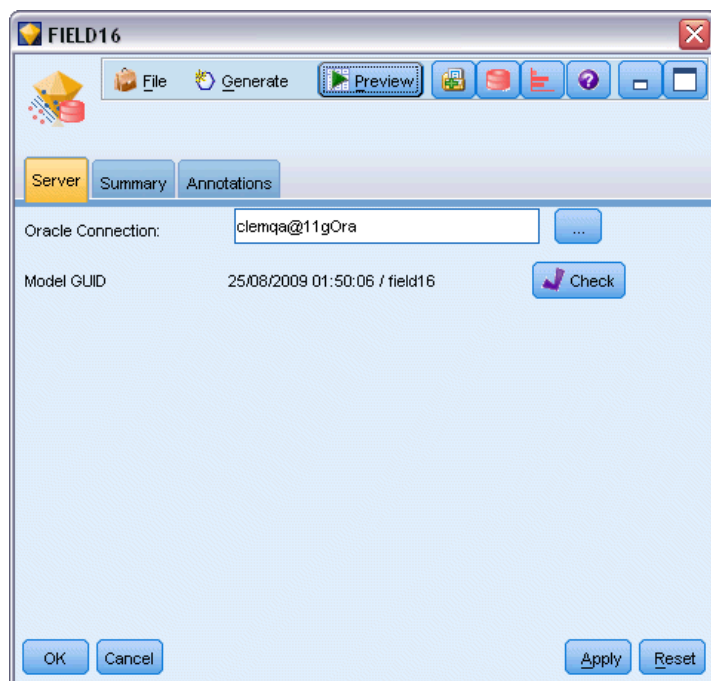
### ***Nugget de modèle Oracle - Onglet Serveur***

Générer un modèle ODM via IBM® SPSS® Modeler crée un modèle dans SPSS Modeler et génère ou remplace un modèle dans la base de données Oracle. Un modèle SPSS Modeler de ce type fait référence au contenu d'un modèle de base de données stocké sur un serveur de base de données. SPSS Modeler peut vérifier la cohérence en stockant une chaîne de **clé de modèle** généré identique dans les modèles SPSS Modeler et Oracle.

La chaîne de clé de chaque modèle Oracle apparaît sous la colonne *Informations sur le modèle* de la boîte de dialogue Modèles de liste. La chaîne de clé d'un modèle SPSS Modeler est affichée comme clé de modèle dans l'onglet Serveur d'un modèle SPSS Modeler (lorsqu'il est placé dans un flux).

Le bouton Vérifier disponible dans l'onglet Serveur d'un nugget de modèle permet de vérifier la correspondance des clés des modèles SPSS Modeler et Oracle. Si vous ne trouvez aucun modèle du même nom dans Oracle ou si les clés de modèle ne correspondent pas, le modèle Oracle a été supprimé ou recréé depuis la création du modèle SPSS Modeler.

Figure 4-31  
Nugget de modèle Oracle - Options de l'onglet Serveur



### ***Nugget de modèle Oracle - Onglet Récapitulatif***

L'onglet Récapitulatif d'un nugget de modèle contient des informations sur le modèle lui-même (*Analyse*), sur les champs utilisés dans le modèle (*Champs*), sur les paramètres utilisés pour la construction du modèle (*Créer des paramètres*), ainsi que sur l'apprentissage du modèle (*Récapitulatif de l'apprentissage*).

Lorsque vous accédez au noeud pour la première fois, l'arborescence des résultats de l'onglet Récapitulatif est réduite. Pour afficher les résultats qui vous intéressent, utilisez la commande à gauche d'un élément pour le développer ou cliquez sur le bouton Développer tout pour afficher tous les résultats. Pour masquer les résultats lorsque vous avez terminé de les consulter, utilisez la commande de développement pour réduire les résultats voulus ou cliquez sur le bouton Réduire tout pour réduire tous les résultats.

**Analyse.** Affiche des informations sur le modèle en question. Si vous avez exécuté un noeud Analyse relié à ce nugget de modèle, les informations issues de l'analyse figureront également dans cette section. [Pour plus d'informations, reportez-vous à la section Noeud Analyse dans le chapitre 6 dans \*Noeuds source, exécution et de sortie de IBM SPSS Modeler 15\*.](#)

**Champs.** Répertoire les champs utilisés comme cibles et entrées lors de la création du modèle.

**Créer des paramètres.** Contient des informations sur les paramètres utilisés lors de la création du modèle.

**Récapitulatif de l'apprentissage.** Indique le type du modèle, le flux utilisé pour le créer, l'utilisateur qui l'a créé, ainsi que sa date et sa durée de création.



## ***Nugget de modèle Oracle - Onglet Paramètres***

L'onglet Paramètres dans le nugget de modèle vous permet de remplacer le paramètre de certaines options du noeud de modélisation pour le scoring.

### ***Arbre décision Oracle***

**Utiliser des coûts de mauvaise réaffectation.** Détermine l'utilisation des coûts de mauvaise réaffectation dans le modèle d'Arbre décision Oracle. [Pour plus d'informations, reportez-vous à la section Coûts de mauvaise réaffectation sur p. 61.](#)

**Identificateur de règle.** S'il est sélectionné (coché), ajoute une colonne d'identificateur de règle au modèle d'Arbre décision Oracle. L'identificateur de règle identifie le noeud dans l'arbre auquel une division spécifique a lieu.

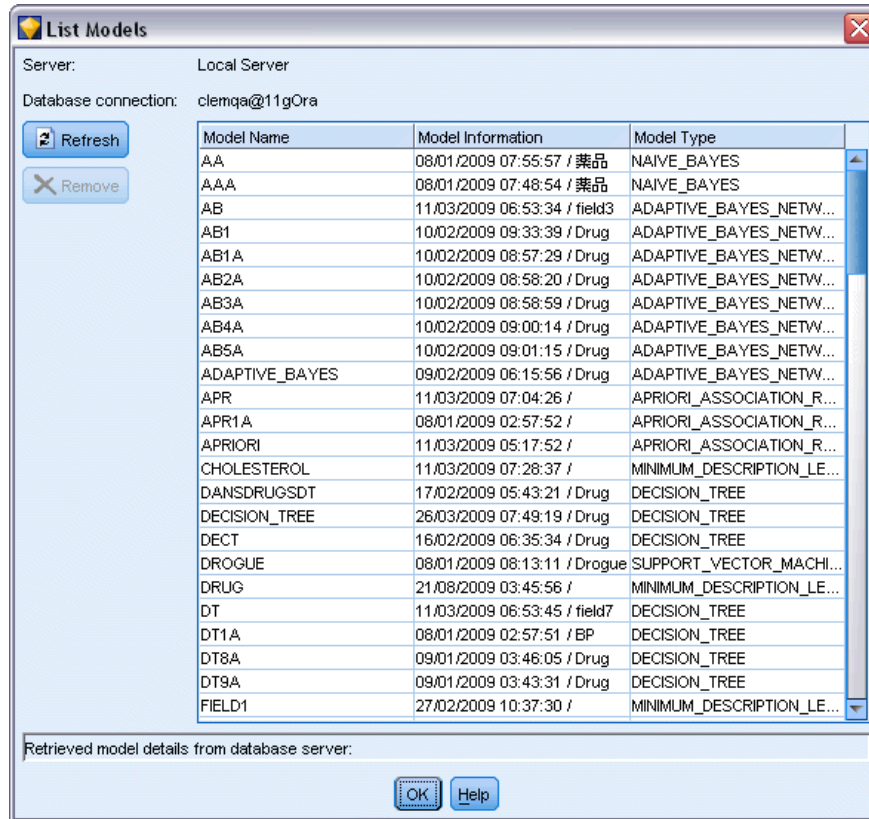
### ***NMF Oracle***

**Afficher toutes les fonctions.** Si sélectionné (coché), affiche l'ID et la confiance de toutes les fonctions, et non pas uniquement les valeurs de la meilleure fonction dans le modèle NMF Oracle.

## ***Liste des modèles Oracle***

Le bouton Répertorier les modèles Oracle Data Mining affiche une boîte de dialogue répertoriant les modèles de base de données existants et permettant d'en supprimer. Vous pouvez afficher cette boîte de dialogue à partir de celle intitulée Programmes externes, et depuis les boîtes de dialogue de création, d'exploration et d'application des noeuds ODM.

Figure 4-32  
Boîte de dialogue de liste des modèles Oracle



Les informations affichées pour chaque modèle sont les suivantes :

- **Nom du modèle.** Nom du modèle, utilisé pour trier la liste
- **Informations sur le modèle.** Informations relatives à la clé du modèle, à savoir les date et heure de création, ainsi que le nom de la colonne cible
- **Type de modèle.** Nom de l'algorithme qui a généré ce modèle

## Oracle Data Miner

Oracle Data Miner est l'interface utilisateur d'Oracle Data Mining (ODM) et remplace l'interface utilisateur IBM® SPSS® Modeler précédente d'ODM. L'interface Oracle Data Miner est conçue pour aider l'analyste à utiliser les algorithmes ODM de manière adéquate. Ces objectifs sont obtenus grâce à différentes méthodes :

- Les utilisateurs requièrent plus d'assistance pour appliquer une méthodologie de préparation des données et de sélection d'algorithme. Oracle Data Miner répond à ce besoin en fournissant des activités de Data mining qui accompagnent les utilisateurs tout au long de la méthodologie appropriée.
- Oracle Data Miner intègre des méthodes heuristiques améliorées et étendues dans les assistants de création et de transformation de modèles pour limiter les risques d'erreur lors de la définition des paramètres de modèle et de transformation.

### Définition d'une connexion Oracle Data Miner

- ▶ Vous pouvez lancer Oracle Data Miner à partir de toutes les versions Oracle, de tous les noeuds d'application et de toutes les boîtes de dialogue de sortie en cliquant sur le bouton **Lancer Oracle Data Miner**.

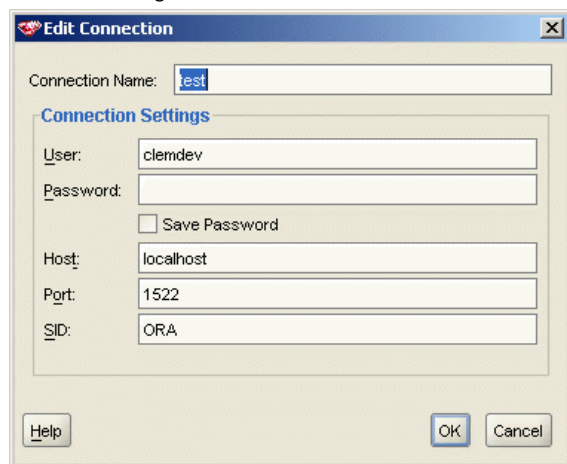
Figure 4-33  
Bouton Lancer Oracle Data Miner



- ▶ La boîte de dialogue **Modifier la connexion** d'Oracle Data Miner apparaît avant que l'application externe Oracle Data Miner soit lancée (à condition que l'option Programmes externes soit correctement définie).

*Remarque* : cette boîte de dialogue n'apparaît que si aucun nom de connexion n'est défini.

Figure 4-34  
Boîte de dialogue Modifier la connexion d'Oracle Data Miner

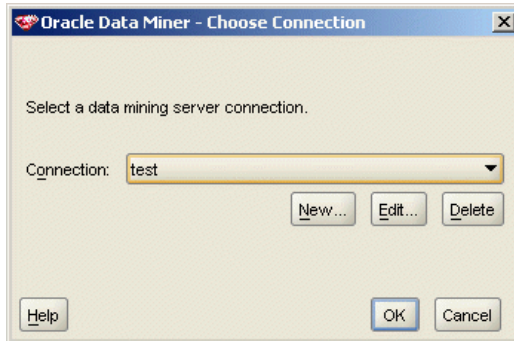


- Saisissez un nom de connexion Data Miner et entrez les informations de serveur Oracle 10gR1 ou 10gR2 appropriées. Le serveur Oracle doit être le serveur indiqué dans SPSS Modeler.

- La boîte de dialogue **Sélectionner une connexion** d'Oracle Data Miner fournit des options pour définir quel nom de connexion, déterminé à l'étape précédente, est utilisé.

Figure 4-35

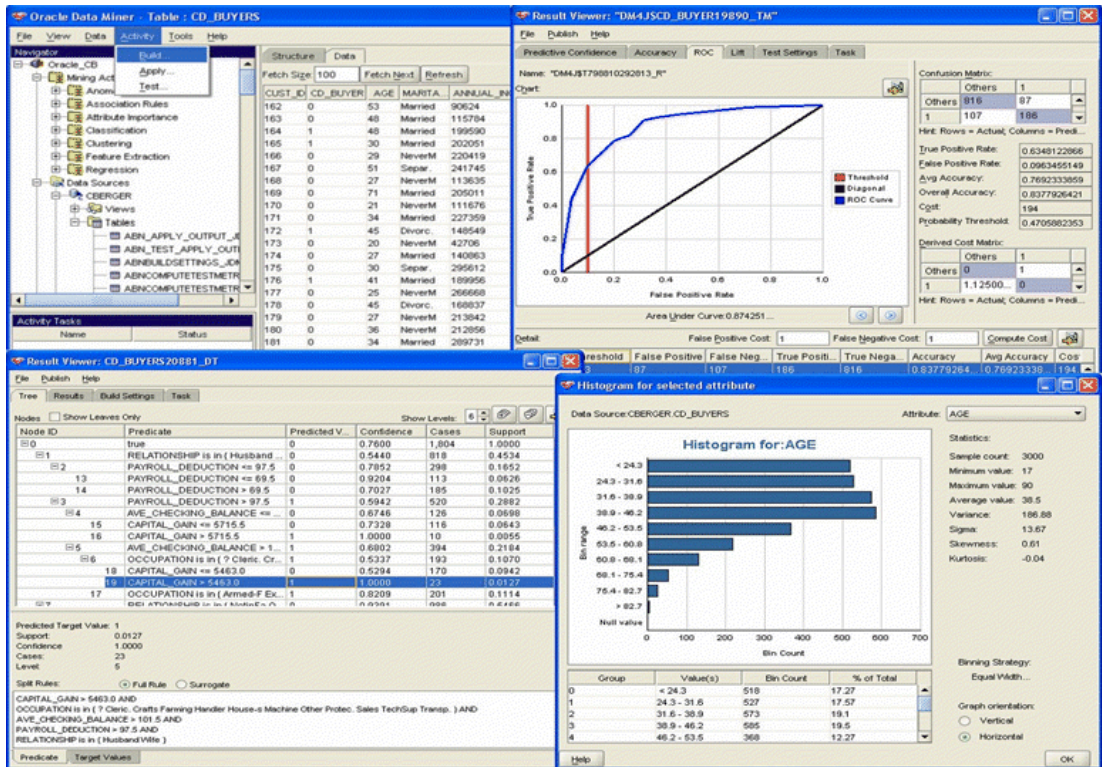
Boîte de dialogue *Sélectionner une connexion* d'Oracle Data Miner



Pour plus d'informations sur la configuration requise, l'installation et l'utilisation d'Oracle Data Miner, reportez-vous à l'article relatif à [Oracle Data Miner](http://www.oracle.com/technology/products/bi/odm/odminer/odminer_install_102.htm) ([http://www.oracle.com/technology/products/bi/odm/odminer/odminer\\_install\\_102.htm](http://www.oracle.com/technology/products/bi/odm/odminer/odminer_install_102.htm)), sur le site Web d'Oracle.

Figure 4-36

Interface utilisateur d'Oracle Data Miner



## Préparation des données

Deux types de préparation des données peuvent s'avérer utiles si vous utilisez les modèles Naive Bayes, Adaptive Bayes et Support Vector Machine fournis avec les algorithmes Oracle Data Mining durant la modélisation :

- **Discrétiser**, conversion des champs d'intervalle numérique continu en catégories pour les algorithmes ne pouvant pas accepter de données continues.
- **Méthode de normalisation**, transformations appliquées à des intervalles numériques pour qu'ils aient des moyennes et des écarts-types similaires.

### **Discrétiser**

Le noeud Discrétiser de IBM® SPSS® Modeler offre de nombreuses méthodes pour effectuer des opérations de création d'intervalles. Une opération de discrétisation est définie comme pouvant être appliquée à un ou plusieurs champs. Exécuter cette opération de discrétisation sur un ensemble de données crée des seuils et permet de générer un noeud Calculer SPSS Modeler. L'opération de calcul peut faire l'objet d'une conversion SQL et être appliquée avant la création et le scoring d'un modèle. Cette approche génère une dépendance entre le modèle et le noeud Calculer qui procède à la discrétisation, mais permet à plusieurs tâches de modélisation de réutiliser les spécifications de discrétisation.

### **Normalisation**

Les champs continus (intervalle numérique) utilisés comme entrées des modèles SVM doivent être normalisés avant la création d'un modèle. Dans le cas de modèles de régression, la normalisation doit également être inversée pour régénérer le score à partir de la sortie du modèle. Les paramètres de modèle SVM vous permettent de sélectionner Centrer-réduire, Min-Max ou Aucun. Oracle crée les coefficients de normalisation dans le cadre du processus de création du modèle. En outre, ces coefficients sont envoyés à SPSS Modeler et stockés avec le modèle. Lors de l'application, ces coefficients sont convertis en formules de calcul SPSS Modeler et utilisés pour préparer les données en vue du scoring, avant de les transmettre au modèle. Dans ce cas, la normalisation est étroitement associée à la tâche de modélisation.

## Oracle Data Mining - Exemples

Plusieurs flux d'échantillons sont fournis pour expliquer l'utilisation d'ODM avec IBM® SPSS® Modeler. Ces flux se trouvent dans le dossier d'installation de SPSS Modeler, sous `\Demos\Database_Modelling\Oracle Data Mining\`.

*Remarque* : Le dossier Démonstrations est accessible à partir du groupe de programmes SPSS Modeler du menu Démarrer de Windows.

Les flux suivants peuvent être utilisés en séquence comme exemple de processus d'exploration de base de données, via l'algorithme SVM fourni avec Oracle Data Mining :

Flux Clementine	Description
<i>1_upload_data.str</i>	Utilisé pour nettoyer et envoyer des données à partir d'un fichier plat vers la base de données.
<i>2_explore_data.str</i>	Offre un exemple d'exploration des données à l'aide de SPSS Modeler
<i>3_build_model.str</i>	Crée le modèle à l'aide de l'algorithme natif de base de données.
<i>4_evaluate_model.str</i>	Utilisé comme exemple d'évaluation de modèle avec SPSS Modeler
<i>5_deploy_model.str</i>	Permet de déployer le modèle de détermination des scores dans la base de données.

*Remarque* : Pour le bon déroulement de l'exemple, vous devez exécuter les flux dans l'ordre. En outre, vous devez mettre à jour les noeuds source et de modélisation de chaque flux de manière à référencer une source de données correcte pour la base de données à utiliser.

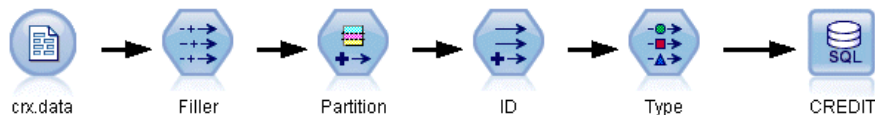
L'ensemble de données utilisé dans les exemples de flux concerne les applications pour carte de crédit et présente un problème de classification avec un mélange de valeurs prédites catégorielles et continues. Pour plus d'informations sur cet ensemble de données, reportez-vous au fichier *crx.names*, qui figure dans le dossier des exemples de flux.

Cet ensemble de données est disponible à partir du référentiel d'apprentissage automatique UCI, sur le site suivant : <ftp://ftp.ics.uci.edu/pub/machine-learning-databases/credit-screening/>.

### Exemple de flux : Envoi de données

Le premier exemple de flux, *1\_upload\_data.str*, est utilisé pour nettoyer et envoyer les données à partir d'un fichier plat vers Oracle.

Figure 4-37  
Exemple de flux utilisé pour envoyer les données



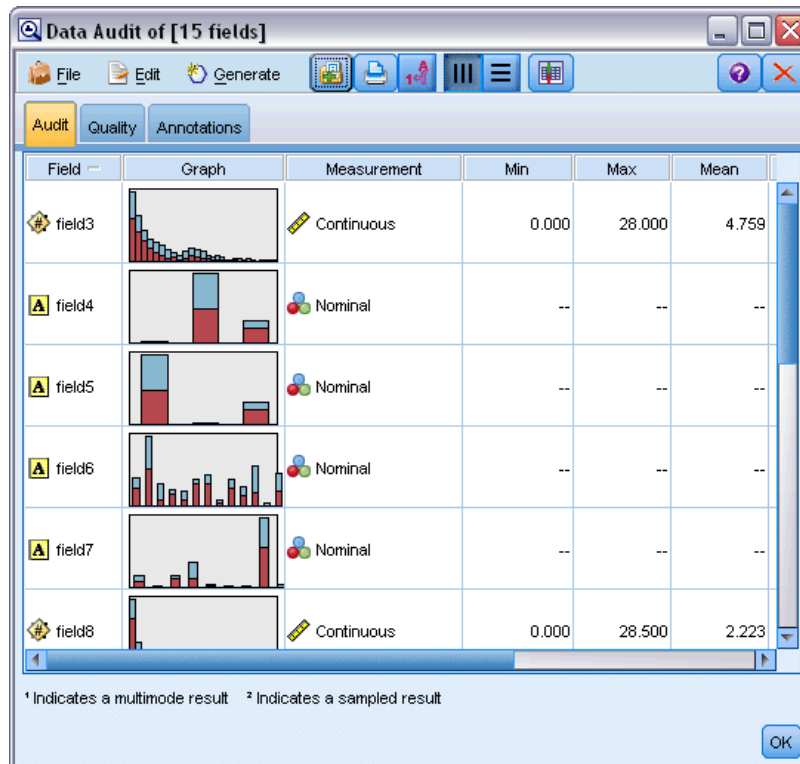
Oracle Data Mining nécessitant un champ d'ID unique ; ce flux initial utilise un noeud Calculer pour ajouter un nouveau champ à l'ensemble de données, *ID*, doté des valeurs uniques 1,2,3, via la fonction IBM® SPSS® Modeler@INDEX.

Le noeud Remplacer est utilisé pour le traitement des valeurs manquantes et remplace les champs vides lus à partir du fichier texte *crx.data* par des valeurs *nulles*.

## Exemple de flux : Exploration des données

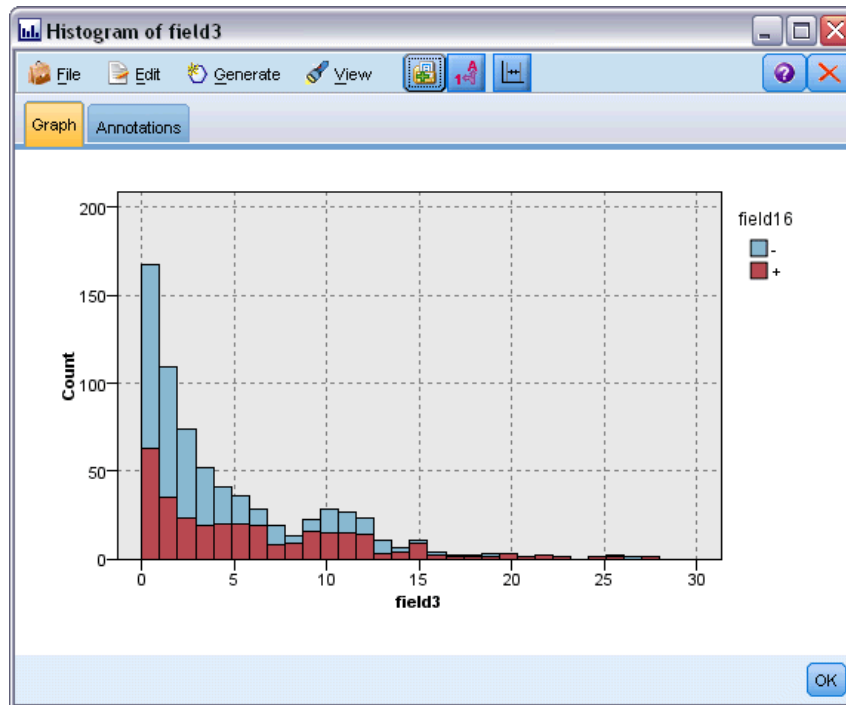
Le deuxième exemple de flux, *2\_explore\_data.str*, est utilisé pour décrire le noeud Audit données pour obtenir un aperçu général des données, y compris les statistiques récapitulatives et les graphiques. [Pour plus d'informations, reportez-vous à la section Noeud Audit données dans le chapitre 6 dans \*Noeuds source, exécution et de sortie de IBM SPSS Modeler 15\*.](#)

Figure 4-38  
Résultats de l'audit de données



Pour obtenir un graphique plus détaillé permettant d'explorer un champ de manière plus précise, double-cliquez sur le graphique voulu dans le rapport Audit données.

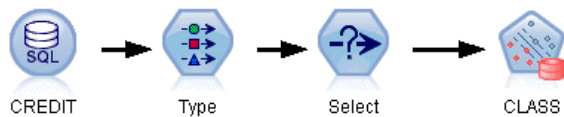
Figure 4-39  
Histogramme créé en double-cliquant sur un graphique dans la fenêtre Audit données



### Exemple de flux : Créer un modèle

Le troisième exemple de flux, `3_build_model.str`, illustre la création de modèles dans IBM® SPSS® Modeler. Effectuez un double-clic sur le noeud source Base de données (intitulé CREDIT) pour définir la source des données. Pour définir les paramètres de création, effectuez un double clic sur le noeud de création (intitulé CLASS à l'origine, puis nommé FIELD16 lorsque la source des données est spécifiée).

Figure 4-40  
Exemple de flux de modélisation de base de données

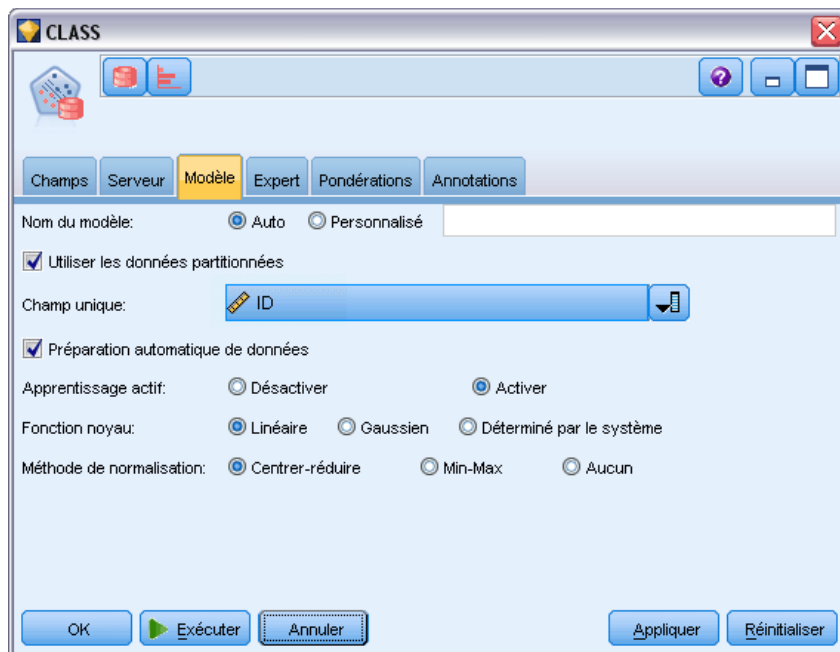


Dans l'onglet Modèle de la boîte de dialogue :

- ▶ Assurez-vous que le champ ID est sélectionné comme champ unique.
- ▶ Assurez-vous que la fonction noyau Linéaire est sélectionnée ainsi que la méthode de normalisation Centrer-réduire.



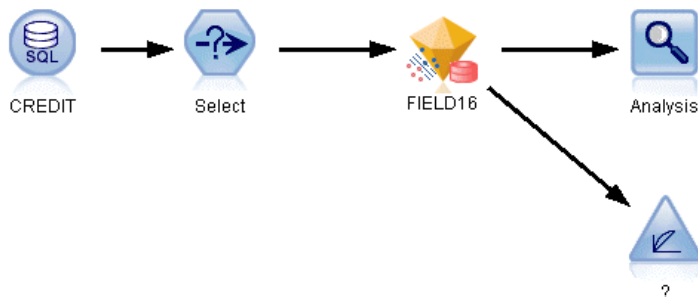
Figure 4-41  
Options de modèle Oracle SVM



### Exemple de flux : Evaluation d'un modèle

Le quatrième exemple de flux, `4_evaluate_model.str`, illustre les avantages que présente l'utilisation de IBM® SPSS® Modeler pour la modélisation dans la base de données. Une fois le modèle exécuté, vous pouvez l'ajouter à nouveau à votre flux de données et l'évaluer à l'aide des divers outils proposés par SPSS Modeler.

Figure 4-42  
Exemple de flux utilisé pour l'évaluation du modèle



### Affichage des résultats d'une modélisation

Reliez un noeud Table au nœud de modèle pour explorer vos résultats. Le champ \$O-field16 affiche les valeurs prédites pour *field16* dans chaque cas et le champ \$OC-field16 affiche la valeur de confiance de cette prévision.

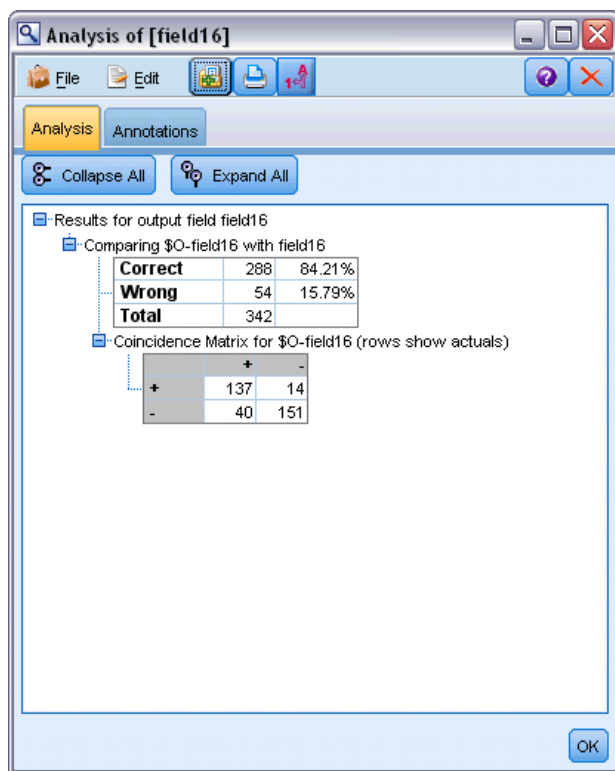
Figure 4-43  
Tableau contenant des informations sur les prédictions générées

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

### ***Evaluation des résultats d'un modèle***

Vous pouvez utiliser un noeud Analyse pour créer une matrice de coïncidence illustrant le motif des correspondances entre chaque champ prédit et son champ cible. Exécutez le noeud Analyse pour afficher les résultats.

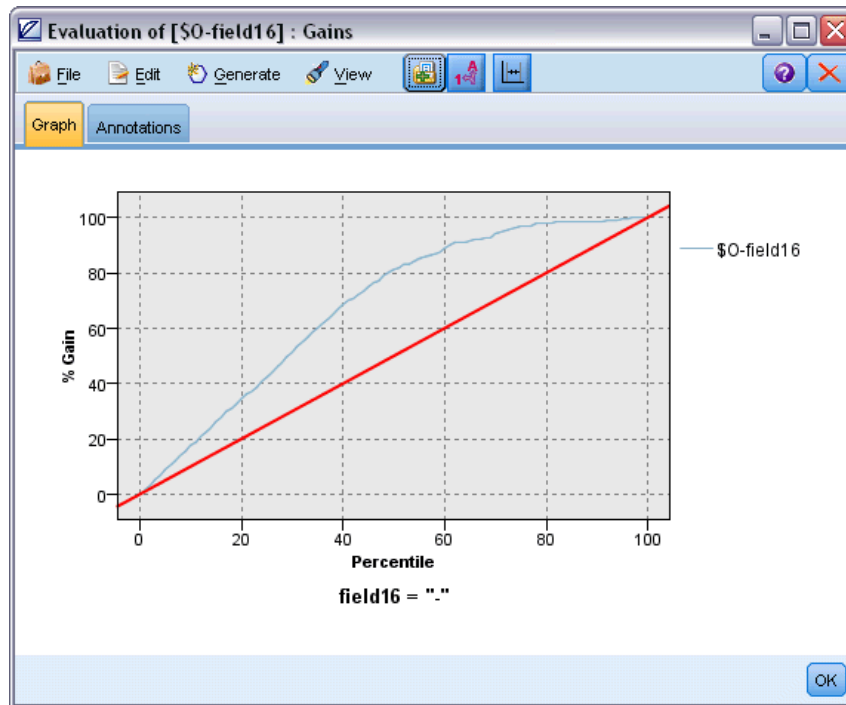
Figure 4-44  
Onglet Analyse comportant des informations sur les résultats d'analyse



Le tableau indique que 84,21% des prévisions créées par l'algorithme Oracle SVM sont correctes.

Vous pouvez utiliser le noeud Evaluation pour créer un graphique de gains qui affiche les améliorations de précision apportées par le modèle. Exécutez le noeud Evaluation pour afficher les résultats.

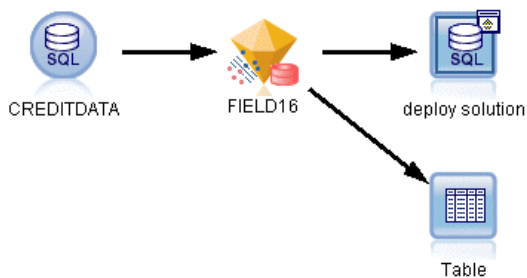
Figure 4-45  
Graphique Gains comportant des informations sur l'amélioration de la précision des modèles



### Exemple de flux : Déploiement d'un modèle

Une fois que la précision du modèle vous convient, vous pouvez le déployer pour l'utiliser avec des applications externes ou pour le republier dans la base de données. Dans le dernier exemple de flux, *5\_deploy\_model.str*, les données sont lues à partir de la table CREDITDATA, puis évaluées et publiées dans la table CREDITSCORES via le noeud Publisher de *solution de déploiement*.

Figure 4-46  
Exemple de flux de modélisation de base de données



Pour plus d'informations, reportez-vous à la section Fonctionnement de IBM SPSS Modeler Solution Publisher dans le chapitre 2 dans *IBM SPSS Modeler 15 Solution Publisher*.

# ***Modélisation de bases de données à l'aide d'IBM InfoSphere Warehouse***

## ***IBM InfoSphere Warehouse et IBM SPSS Modeler***

IBM InfoSphere Warehouse (ISW) fournit un ensemble d'algorithmes de Data mining incorporés dans le SGBDR du DB2 d'IBM. IBM® SPSS® Modeler prend en charge des noeuds qui prennent en charge l'intégration des algorithmes d'IBM :

- Arbres décision
- Règles d'association
- Classification démographique
- Classification non supervisée Kohonen
- Règles de séquence
- Régression de transformation
- Régression linéaire
- Régression polynomiale
- Naive Bayes
- Régression logistique
- Séries temporelles

Pour plus d'informations sur ces algorithmes, reportez-vous à la documentation fournie avec votre installation d'IBM InfoSphere Warehouse.

## ***Conditions requises pour l'intégration à IBM InfoSphere Warehouse***

Les conditions préalables suivantes sont obligatoires pour réaliser une modélisation dans la base de données via InfoSphere Warehouse Data Mining. N'hésitez pas à contacter l'administrateur de base de données pour vous assurer que ces conditions sont remplies.

- IBM® SPSS® Modeler en cours d'exécution en parallèle avec une installation IBM® SPSS® Modeler Server sur Windows ou UNIX.
  - IBM DB2 Data Warehouse Edition Version 9.1
- ou*
- IBM InfoSphere Warehouse Version 9.5 Edition Entreprise
  - Une source de données ODBC pour la connexion à DB2, comme indiqué ci-après.

*Remarque* : La modélisation de base de données et l'optimisation SQL requièrent l'activation de la connectivité à SPSS Modeler Server sur l'ordinateur SPSS Modeler. Avec ce paramètre activé, vous pouvez accéder aux algorithmes de la base de données, effectuer le push back de SQL

directement depuis SPSS Modeler et accéder à SPSS Modeler Server. Pour vérifier l'état actuel de la licence, choisissez ce qui suit dans le menu SPSS Modeler.

Aide > A propos de > Informations supplémentaires

Si la connectivité est activée, vous voyez l'option Activation du serveur dans l'onglet Etat de la licence.

Pour plus d'informations, reportez-vous à la section Connexion au IBM SPSS Modeler Server dans le chapitre 3 dans *Guide de l'utilisateur de IBM SPSS Modeler 15*.

## **Activation de l'intégration à IBM InfoSphere Warehouse**

Pour activer l'intégration de IBM® SPSS® Modeler à IBM InfoSphere Warehouse (ISW) Data Mining, vous devez configurer ISW, créer une source de données ODBC, activer cette intégration dans la boîte de dialogue Programmes externes de SPSS Modeler, puis activer la génération et l'optimisation SQL.

### **Configuration d'ISW**

Pour installer et configurer ISW, suivez les instructions fournies dans le guide d'*installation d'InfoSphere Warehouse*.

### **Création d'une source de données ODBC pour ISW**

Pour activer la connexion entre ISW et SPSS Modeler, vous devez créer un nom de la source de données ODBC (DSN).

Avant de créer un DSN, vous devez avoir des connaissances de base des sources de données et des pilotes ODBC, ainsi que de la prise en charge de la base de données dans SPSS Modeler. [Pour plus d'informations, reportez-vous à la section Accès aux données dans le chapitre 2 dans \*Guide d'administration et des performances de IBM SPSS Modeler Server 15\*](#).

Si IBM® SPSS® Modeler Server et IBM InfoSphere Warehouse Data Mining sont exécutés sur des hôtes différents, créez le même DSN ODBC sur chaque hôte. Veillez à utiliser le même nom de DSN sur chaque hôte.

- ▶ Installez les pilotes ODBC. Ceux-ci sont disponibles sur le disque d'installation IBM® SPSS® Data Access Pack fourni avec cette version. Exécutez le fichier *setup.exe* pour démarrer le programme d'installation et sélectionnez les pilotes appropriés. Suivez les instructions à l'écran pour installer les pilotes.
- ▶ Créez le DSN (nom de source de données).

*Remarque* : La séquence de menus dépend de la version de Windows que vous utilisez.

- **Windows XP.** Dans le menu Démarrer, sélectionnez Panneau de configuration. Double-cliquez sur Outils d'administration, puis sur Sources de données (ODBC).

- **Windows Vista.** Dans le menu Démarrer, sélectionnez Panneau de configuration; puis Maintenance du système. Double-cliquez sur Outils d'administration, sélectionnez Sources de données (ODBC), puis cliquez sur Ouvrir.
- **Windows 7.** Dans le menu Démarrer, sélectionnez Panneau de configuration; puis Sécurité du système, puis Outils d'administration. Sélectionnez Sources de données (ODBC), puis cliquez sur Ouvrir.
- ▶ Cliquez sur l'onglet du DSN système, puis sur Ajouter.
- ▶ Sélectionnez le pilote SPSS OEM 6.0 DB2 Wire Protocol.
- ▶ Cliquez sur Terminer.
- ▶ Dans la boîte de dialogue de configuration du pilote ODBC DB2 Wire Protocol :
  - Indiquez le nom de la source de données.
  - Pour l'adresse IP, indiquez le nom d'hôte du serveur sur lequel est exécuté le SGBDR DB2.
  - Acceptez la valeur par défaut du port TCP (50000).
  - Indiquez le nom de la base de données à laquelle vous devez vous connecter.
- ▶ Cliquez sur Test Connection.
- ▶ Dans la boîte de dialogue Logon to DB2 Wire Protocol, entrez le nom et le mot de passe utilisateur qui vous ont été donnés par l'administrateur de base de données, puis cliquez sur OK.

Le message indiquant que la connexion est établie apparaît.
- ▶ **Pilote ODBC du DB2 d'IBM.** Si votre pilote ODBC est le pilote ODBC du DB2 d'IBM, suivez les étapes ci-après pour créer un DSN ODBC :
  - ▶ Dans l'administrateur de source de données ODBC, cliquez sur l'onglet du DSN système, puis sur Add.
  - ▶ Sélectionnez le pilote ODBC IBM DB2 et cliquez sur Finish.
  - ▶ Dans la fenêtre d'ajout du pilote ODBC du DB2 d'IBM, entrez le nom de la source de données, puis cliquez sur Ajouter pour l'alias de base de données.
  - ▶ Dans l'onglet de source de données de la fenêtre CLI/ODBC Settings—<nom de la source de données>, entrez l'ID et le mot de passe utilisateur attribués par l'administrateur de base de données, puis cliquez sur l'onglet TCP/IP.
  - ▶ Dans l'onglet TCP/IP, entrez les informations suivantes :
    - Le nom de la base de données à laquelle vous souhaitez vous connecter.
    - Un nom d'alias de base de données (8 caractères au maximum).
    - Le nom d'hôte du serveur de base de données auquel vous souhaitez vous connecter.
    - Le numéro de port de la connexion.
  - ▶ Dans l'onglet Security Options, sélectionnez Specify the security options (Optional), puis acceptez la valeur par défaut (Use authentication value in server's DBM Configuration).
  - ▶ Dans l'onglet Data Source, cliquez sur Connect.

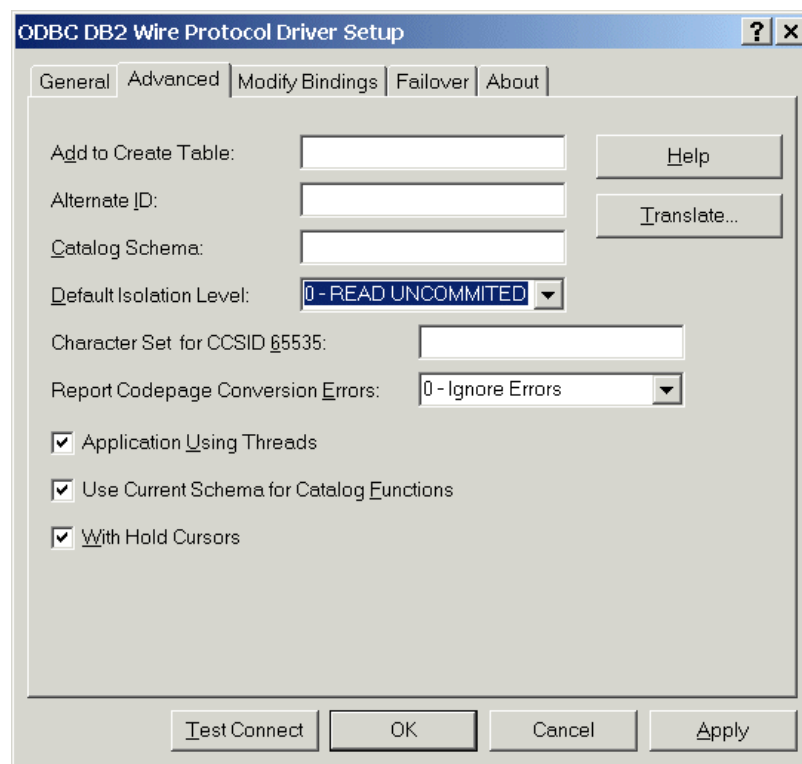
Le message qui apparaît indique le test de connexion a réussi.

### **Configuration d'ODBC pour les commentaires (facultatif)**

Pour recevoir des commentaires de IBM InfoSphere Warehouse Data Mining lors de la création de modèles et pour autoriser SPSS Modeler à annuler cette création, suivez la procédure ci-après pour configurer la source de données ODBC créée dans la section précédente. Cette étape de configuration autorise SPSS Modeler à lire les données de DB2 non validées auprès de la base de données en exécutant les transactions de manière simultanée. Si vous avez des doutes quant aux implications de cette modification, consultez l'administrateur de base de données.

Figure 5-1

Boîte de dialogue Installation du pilote ODBC DB2 Wire Protocol, onglet Options avancées

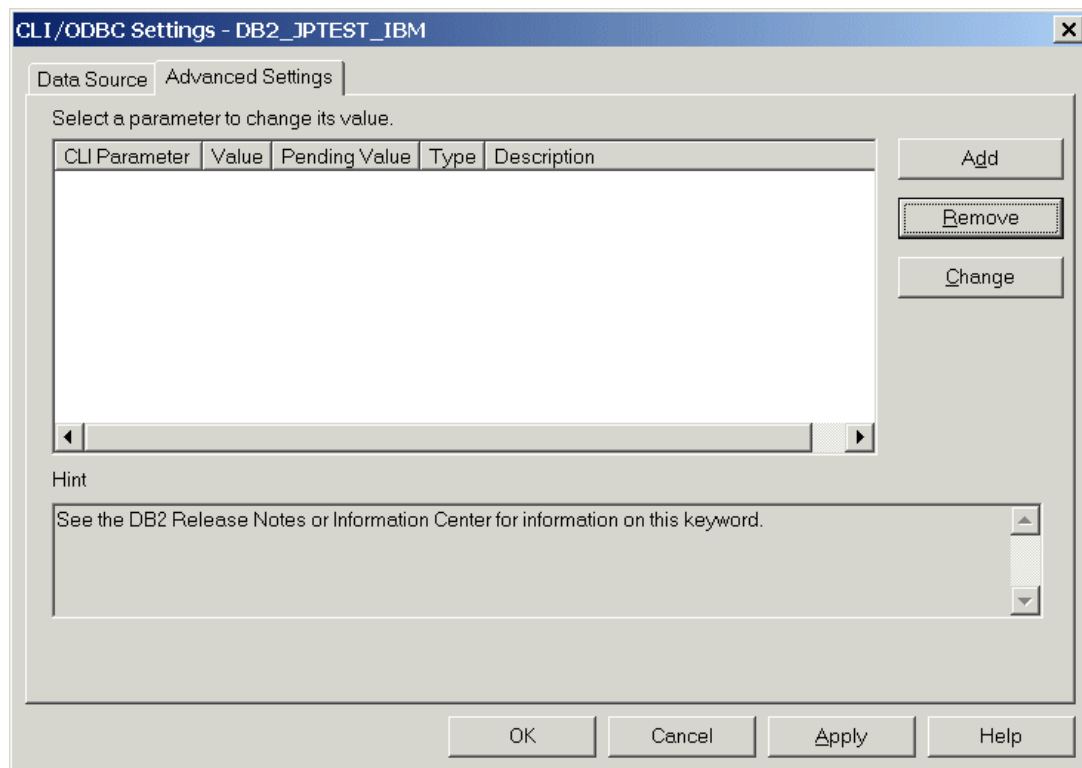


**Pilote SPSS OEM 6.0 DB2 Wire Protocol.** Pour le pilote Connect ODBC, procédez comme suit :

- ▶ Démarrez l'administrateur de source de données ODBC, sélectionnez la source de données créée dans la section précédente et cliquez sur le bouton Configurer.
- ▶ Dans la boîte de dialogue ODBC DB2 Wire Protocol Driver Setup, cliquez sur l'onglet Advanced.
- ▶ Définissez le niveau d'isolation par défaut sur 0-READ UNCOMMITTED, puis cliquez sur OK.



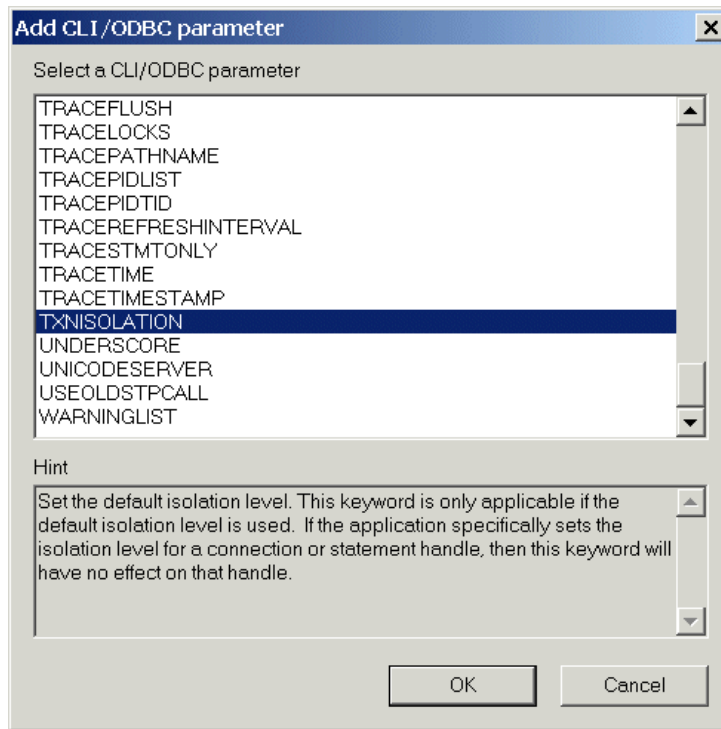
Figure 5-2  
Boîte de dialogue CLI/ODBC Settings, onglet Advanced Settings



**Pilote ODBC du DB2 d'IBM.** Pour le pilote DB2 d'IBM, procédez comme suit :

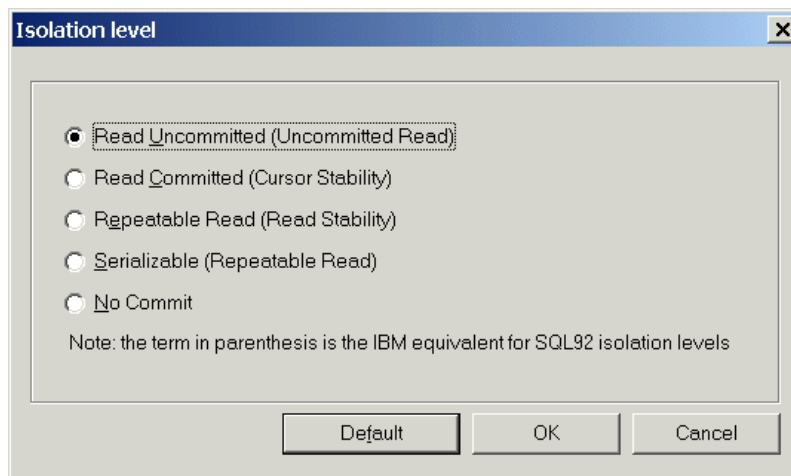
- ▶ Démarrez l'administrateur de source de données ODBC, sélectionnez la source de données créée dans la section précédente, puis cliquez sur le bouton Configurer.
- ▶ Dans la boîte de dialogue CLI/ODBC Settings, cliquez sur l'onglet Advanced Settings, puis sur le bouton Add.

**Figure 5-3**  
*Boîte de dialogue CLI/ODBC Parameter*



- Dans la boîte de dialogue Add CLI/ODBC Parameter, sélectionnez le paramètre TXNISOLATION, puis cliquez sur OK.

**Figure 5-4**  
*Boîte de dialogue Isolation level*



- Dans la boîte de dialogue Isolation level, sélectionnez Read Uncommitted, puis cliquez sur OK.
- Dans la boîte de dialogue CLI/ODBC Settings, cliquez sur OK pour terminer la configuration.

Les commentaires rapportés par IBM InfoSphere Warehouse Data Mining apparaissent au format suivant :

<ITERATIONNO> / <PROGRESS> / <KERNELPHASE>

où :

- <ITERATIONNO> indique le numéro du passage actuel sur les données, à partir de 1.
- <PROGRESS> indique la progression de l'itération en cours sous la forme d'un chiffre compris entre 0,0 et 1,0.
- <KERNELPHASE> décrit la phase en cours de l'algorithme d'exploration.

### **Activation de l'intégration de IBM InfoSphere Warehouse Data Mining dans IBM SPSS Modeler**

Pour permettre à SPSS Modeler d'utiliser DB2 avec IBM InfoSphere Warehouse Data Mining, vous devez fournir au préalable quelques spécifications dans la boîte de dialogue Programmes externes.

- ▶ Dans les menus SPSS Modeler, sélectionnez :  
Outils > Options > Programmes externes
- ▶ Cliquez sur l'onglet IBM InfoSphere Warehouse.

**Activation de l'intégration d'InfoSphere Warehouse Data Mining.** Active la palette Modélisation de la base de données (si elle n'est pas déjà visible) en bas de la fenêtre SPSS Modeler et ajoute les noeuds des algorithmes ISW Data Mining.

**Connexion DB2.** Indique la source de données ODBC DB2 par défaut utilisée pour créer et stocker les modèles. Ce paramètre peut être ignoré pour la création de modèle individuelle et les noeuds de modèle générés. Cliquez sur le bouton ... pour choisir la source de données.

La connexion de base de données utilisée pour la modélisation peut être identique à celle utilisée pour accéder aux données. Par exemple, un flux peut accéder aux données à partir d'une base de données DB2, envoyer les données à SPSS Modeler pour le nettoyage ou toute autre manipulation, puis télécharger ces données vers une autre base de données DB2 à des fins de modélisation. Sinon, au cas où les données d'origine se trouvent dans un fichier plat ou une autre source (non-DB2), vous devez les envoyer à DB2 pour la modélisation. Dans tous les cas, les données sont automatiquement envoyées vers une table temporaire créée dans la base de données utilisée pour la modélisation, si nécessaire.

### **Avertir au moment de remplacer un modèle d'intégration InfoSphere Warehouse Data Mining.**

Sélectionnez cette option pour vous assurer que les modèles stockés dans la base de données ne sont pas ignorés par SPSS Modeler sans que vous en soyez informé.

**Répertoire des modèles d'InfoSphere Warehouse Data Mining.** Cette option vous permet de répertorier et de supprimer les modèles stockés dans DB2. [Pour plus d'informations, reportez-vous à la section Liste des modèles de base de données sur p. 118.](#)

**Activation de l'exécution d'InfoSphere Warehouse Data Mining Visualization.** Si vous avez installé le module de visualisation de DB2, vous devez l'activer ici pour que SPSS Modeler puisse l'utiliser.

**Chemin vers l'exécutable de visualisation.** L'emplacement du fichier exécutable du module Visualisation (s'il est installé), par exemple *C:\Program Files\IBM\ISWarehouse\Im\IMVisualization\bin\imvisualizer.exe*.

**Répertoire du plug-in de visualisation de séries temporelles.** L'emplacement du plug-in flash du module Visualisation de séries temporelles (s'il est installé), par exemple *C:\Program Files\IBM\ISWShared\plugins\com.ibm.datatools.datamining.imvisualization.flash\_2.2.1.v20091111\_0915*.

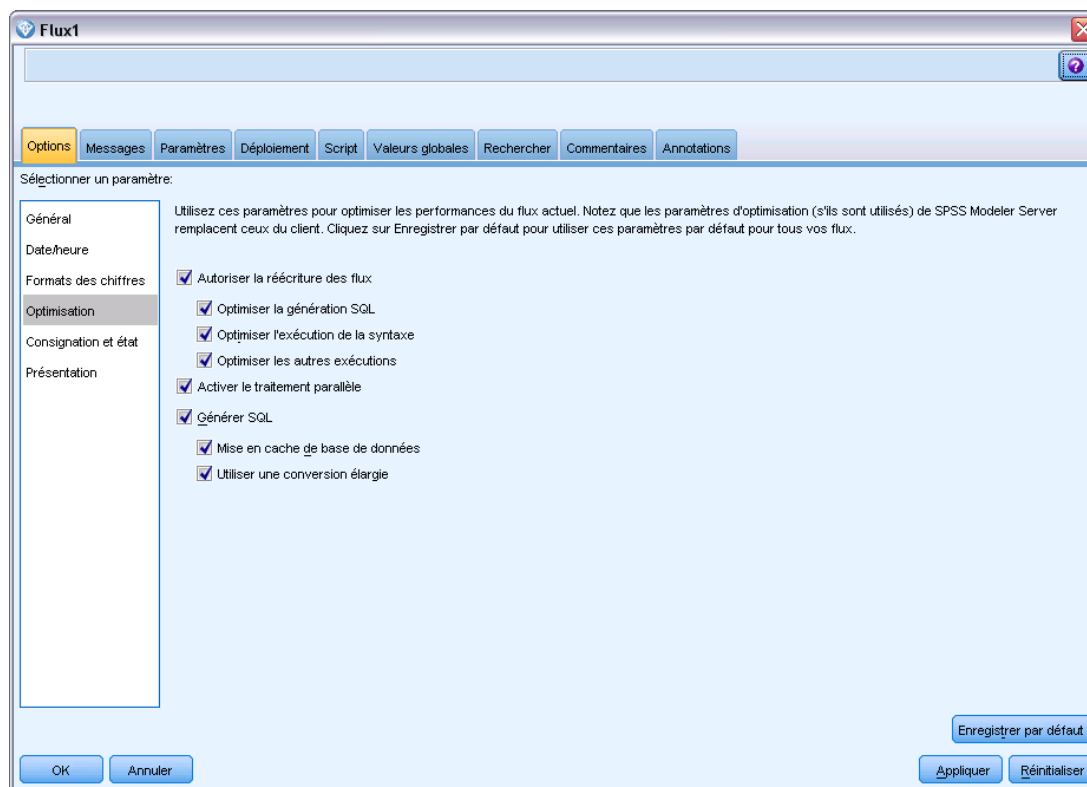
**Activation des options de puissance d'InfoSphere Warehouse Data Mining.** Vous pouvez définir une limite de consommation de la mémoire sur un algorithme d'exploration dans la base de données et indiquer d'autres options arbitraires sous forme de lignes de commande pour des modèles spécifiques. La limite de mémoire permet de contrôler la consommation de mémoire et d'indiquer une valeur pour l'option de puissance `-buf`. Vous pouvez indiquer d'autres options de puissance ici sous la forme de lignes de commande, qui sont ensuite passées à IBM InfoSphere Warehouse Data Mining. [Pour plus d'informations, reportez-vous à la section Options de puissance sur p. 121.](#)

**Vérification de la version d'InfoSphere Warehouse.** Vérifie la version d'IBM InfoSphere Warehouse utilisée et signale une erreur si vous essayez d'utiliser une fonction de Data Mining non prise en charge par votre version.

#### **Activation de la génération SQL et de l'optimisation**

- ▶ Dans les menus SPSS Modeler, sélectionnez :  
Outils > Propriétés du flux > Options

Figure 5-5  
Paramètres d'optimisation



- ▶ Cliquez sur l'option Optimisation dans le volet de navigation.
- ▶ Confirmez que l'option Générer SQL est bien activée. Ce paramètre doit être utilisé pour que la modélisation de base de données puisse fonctionner.
- ▶ Sélectionnez Optimiser la génération SQL et Optimiser les autres exécutions (cette opération n'est pas forcément nécessaire, mais elle est vivement recommandée pour optimiser les performances).

Pour plus d'informations, reportez-vous à la section Définition des options d'optimisation pour les flux dans le chapitre 5 dans *Guide de l'utilisateur de IBM SPSS Modeler 15*.

## Création de modèles à l'aide de IBM InfoSphere Warehouse Data Mining

La création de modèles avec IBM InfoSphere Warehouse Data Mining requiert la présence de l'ensemble de données d'apprentissage dans une table ou une vue de la base de données DB2. Si les données ne se trouvent pas dans DB2 ou doivent être traitées dans IBM® SPSS® Modeler, dans le cadre de la préparation des données ne pouvant pas être exécutées dans DB2, les données sont automatiquement envoyées vers une table temporaire dans DB2 avant la création du modèle.

## Détermination de scores et déploiement du modèle

La détermination de scores du modèle intervient toujours dans DB2 et est systématiquement exécutée par IBM InfoSphere Warehouse Data Mining. Vous devrez peut-être envoyer l'ensemble de données à une table temporaire si les données proviennent de, ou doivent être préparées dans IBM® SPSS® Modeler. Pour l'Arbre décision, la régression et les modèles de classification dans SPSS Modeler, en général, une seule prévision et la probabilité ou confiance associées sont livrées. De plus, une option utilisateur permettant d'afficher les confiances pour chaque revenu possible (semblable à celles trouvées dans la régression logistique) correspond à une option de temps de score disponible dans l'onglet Paramètres du nugget de modèle (case à cocher Inclure les confiances de toutes les classes). Pour les modèles de séquence et d'association dans SPSS Modeler, différentes valeurs sont fournies. SPSS Modeler peut évaluer les modèles d'IBM InfoSphere Warehouse Data Mining à partir de flux publiés pour être exécutés via IBM® SPSS® Modeler Solution Publisher.

Les champs suivants sont générés par les modèles de scoring :

Table 5-1  
Champs de scoring des modèles

Type de modèle	Colonnes de score	Signification
Arbres décision	\$I- <i>champ</i>	Meilleure prévision pour le <i>champ</i> .
	\$IC- <i>champ</i>	Confiance de la meilleure prévision pour le <i>champ</i> .
	\$IC- <i>value1</i> , ..., \$IC- <i>valueN</i>	(facultatif) La confiance de chacune des <i>N</i> valeurs possibles pour le <i>champ</i> .
Régression	\$I- <i>champ</i>	Meilleure prévision pour le <i>champ</i> .
	\$IC- <i>champ</i>	Confiance de la meilleure prévision pour le <i>champ</i> .
Classification	\$I- <i>nom_du_modèle</i>	Meilleure affectation de classe pour l'enregistrement d'entrée.
	\$IC- <i>nom_du_modèle</i>	Confiance de la meilleure affectation de classe pour l'enregistrement d'entrée.
Association	\$I- <i>nom_du_modèle</i>	Identificateur de règle correspondante.
	\$IH- <i>nom_du_modèle</i>	Élément d'en-tête.
	\$IHN- <i>nom_du_modèle</i>	Nom de l'élément d'en-tête.
	\$IS- <i>nom_du_modèle</i>	Valeur de prise en charge de règle correspondante.
	\$IC- <i>nom_du_modèle</i>	Valeur de confiance de règle correspondante.
	\$IL- <i>nom_du_modèle</i>	Valeur de Lift (augmentation) de règle correspondante.

Type de modèle	Colonnes de score	Signification
	\$IMB- <i>nom_du_modèle</i>	Le nombre d'éléments du corps correspondant ou des jeux d'éléments du corps (car tous les éléments du corps ou les jeux d'éléments du corps doivent correspondre à ce nombre, égal au nombre d'éléments du corps ou des jeux d'éléments du corps).
Séquence	\$I- <i>nom_du_modèle</i>	Identificateur de règle correspondante
	\$IH- <i>nom_du_modèle</i>	Élément d'en-tête de règle correspondante
	\$IHN- <i>nom_du_modèle</i>	Noms d'éléments dans l'élément d'en-tête de règle correspondante
	\$IS- <i>nom_du_modèle</i>	Valeur de prise en charge de règle correspondante
	\$IC- <i>nom_du_modèle</i>	Valeur de confiance de règle correspondante
	\$IL- <i>nom_du_modèle</i>	Valeur de Lift (augmentation) de règle correspondante
	\$IMB- <i>nom_du_modèle</i>	Le nombre d'éléments du corps correspondant ou des jeux d'éléments du corps (car tous les éléments du corps ou les jeux d'éléments du corps doivent correspondre à ce nombre, égal au nombre d'éléments du corps ou des jeux d'éléments du corps)
Naive Bayes	\$I- <i>champ</i>	Meilleure prévision pour le <i>champ</i> .
	\$IC- <i>champ</i>	Confiance de la meilleure prévision pour le <i>champ</i> .
Régression logistique	\$I- <i>champ</i>	Meilleure prévision pour le <i>champ</i> .
	\$IC- <i>champ</i>	Confiance de la meilleure prévision pour le <i>champ</i> .

## Gestion des modèles DB2

La création d'un modèle IBM InfoSphere Warehouse Data Mining via IBM® SPSS® Modeler crée un modèle dans SPSS Modeler et crée ou remplace un modèle dans la base de données DB2. Le modèle SPSS Modeler de ce type fait référence au contenu d'un modèle de base de données stocké sur un serveur de base de données. SPSS Modeler peut vérifier la cohérence en stockant une chaîne de clé de modèle générée identique dans les modèles SPSS Modeler et DB2.

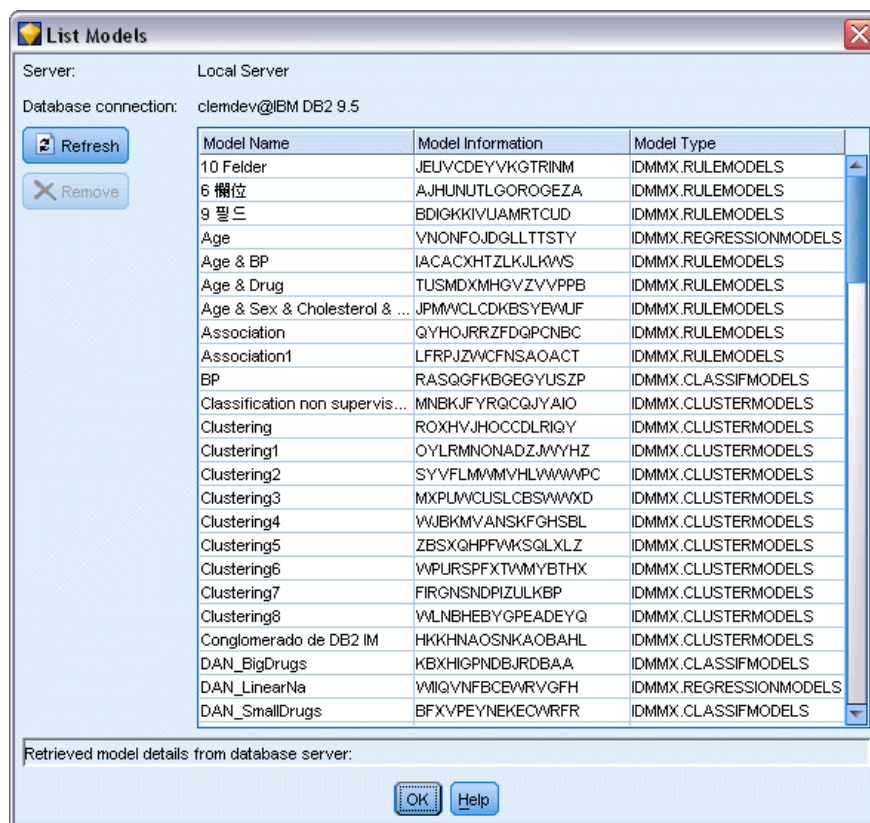
La chaîne de clé de chaque modèle DB2 est affichée dans la colonne d'*informations sur le modèle* de la boîte de dialogue des modèles de base de données. La chaîne de clé d'un modèle SPSS Modeler est affichée comme clé de modèle dans l'onglet Serveur d'un modèle SPSS Modeler (lorsqu'il est placé dans un flux).

Le bouton Vérifier permet de vérifier la correspondance entre les clés du modèle SPSS Modeler et du modèle DB2. Si vous ne trouvez aucun modèle du même nom dans DB2 ou si les clés de modèle ne correspondent pas, le modèle DB2 a été supprimé ou recréé depuis la création du modèle SPSS Modeler. [Pour plus d'informations, reportez-vous à la section Nugget de modèle ISW - Onglet Serveur sur p. 155.](#)

## Liste des modèles de base de données

IBM® SPSS® Modeler propose une boîte de dialogue qui répertorie les modèles stockés dans IBM InfoSphere Warehouse Data Mining et permet de les supprimer.

Figure 5-6  
Boîte de dialogue Modèles de liste de DB2



Cette boîte de dialogue est accessible à partir de la boîte de dialogue des programmes externes IBM et des boîtes de dialogue de création, de navigation et d'application des noeuds liés à IBM InfoSphere Warehouse Data Mining. Les informations affichées pour chaque modèle sont les suivantes :

- Nom de modèle (nom du modèle, utilisé pour trier la liste)



- Informations sur le modèle (informations-clés du modèle, issues d'une clé aléatoire générée lors de la création du modèle par SPSS Modeler)
- Type de modèle (table DB2 dans laquelle IBM InfoSphere Warehouse Data Mining a stocké le modèle).

### **Navigation dans les modèles**

L'outil Visualizer est la seule méthode permettant de parcourir les modèles InfoSphere Warehouse Data Mining. L'outil peut être installé en option avec InfoSphere Warehouse Data Mining. [Pour plus d'informations, reportez-vous à la section Activation de l'intégration à IBM InfoSphere Warehouse sur p. 108.](#)

- Cliquez sur Affichage pour lancer l'outil de visualisation. L'affiche de l'outil dépend du type de noeud généré. L'outil de visualisation peut, par exemple, afficher les classes prédites s'il est lancé à partir d'un nugget de modèle d'arbre de décision ISW.
- Cliquez sur Résultats de test (arbres décision et séquence uniquement) pour lancer l'outil de visualisation et afficher la qualité générale du modèle généré.

### **Exportation de modèles et génération de noeuds**

Vous pouvez réaliser des actions d'importation et d'exportation PMML sur les modèles IBM InfoSphere Warehouse Data Mining. Le PMML exporté est le PMML d'origine généré par IBM InfoSphere Warehouse Data Mining. La fonction d'exportation renvoie le modèle au format PMML.

Vous pouvez exporter un récapitulatif et une structure de modèle dans des fichiers au format texte et HTML. Vous pouvez générer les noeuds Filtrer, Sélectionner et Calculer nécessaires. Pour plus d'informations, reportez-vous au chapitre consacré à l'exportation de modèles dans le *guide de l'utilisateur de IBM® SPSS® Modeler*.

### **Paramètres de noeud communs à tous les algorithmes**

Les paramètres suivants sont communs à de nombreux algorithmes IBM InfoSphere Warehouse Data Mining :

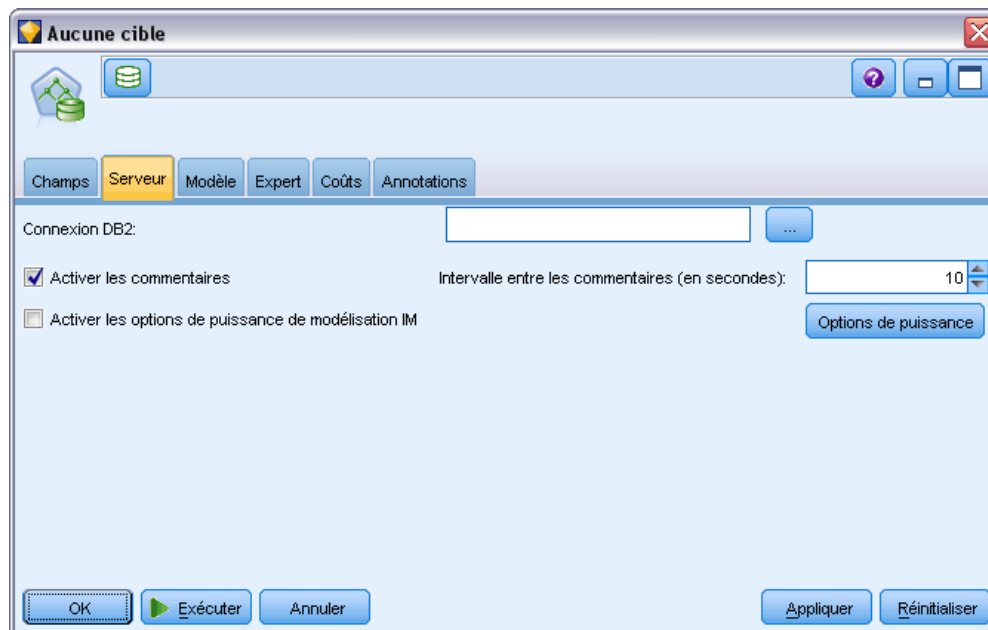
**Cible et variables indépendantes.** Vous pouvez indiquer une cible et des variables indépendantes à l'aide du noeud Typer ou manuellement en utilisant l'onglet Champs du noeud de création du modèle, procédure standard dans IBM® SPSS® Modeler.

**sources de données ODBC.** Cette option vous permet d'ignorer la source de données ODBC par défaut du modèle en cours. (La valeur par défaut est indiquée dans la boîte de dialogue Programmes externes. [Pour plus d'informations, reportez-vous à la section Activation de l'intégration à IBM InfoSphere Warehouse sur p. 108.](#))

### Options de l'onglet du serveur ISW

Vous pouvez indiquer la connexion DB2 utilisée pour envoyer les données à modéliser. Si nécessaire, vous pouvez sélectionner une connexion dans l'onglet Serveur pour chaque noeud de modélisation afin d'ignorer la connexion DB2 par défaut indiquée dans la boîte de dialogue Programmes externes. [Pour plus d'informations, reportez-vous à la section Activation de l'intégration à IBM InfoSphere Warehouse sur p. 108.](#)

Figure 5-7  
Onglet Serveur ISW



La connexion utilisée pour la modélisation peut être identique à la connexion utilisée dans le noeud source d'un flux. Par exemple, un flux peut accéder aux données à partir d'une base de données DB2, envoyer les données à IBM® SPSS® Modeler pour le nettoyage ou toute autre manipulation, puis télécharger ces données vers une autre base de données DB2 à des fins de modélisation.

Le nom de la source de données ODBC est incorporé de manière efficace dans chaque flux SPSS Modeler. Si un flux créé sur un hôte est exécuté sur un autre hôte, le nom de la source de données doit être identique sur chaque hôte. Vous pouvez également sélectionner une autre source de données dans l'onglet Serveur de chaque source ou noeud de modélisation.

Vous pouvez obtenir des commentaires lors de la création d'un modèle avec les options suivantes :

- **Activer les commentaires.** Sélectionnez cette option (désactivée par défaut) pour obtenir des commentaires lors de la création d'un modèle.
- **Intervalle entre les commentaires (en secondes).** Indiquez la fréquence à laquelle SPSS Modeler extrait des commentaires sur la progression de la création du modèle.

**Activation des options de puissance d'InfoSphere Warehouse Data Mining.** Sélectionnez cette option pour activer le bouton Options de puissance, qui vous permet de spécifier plusieurs options avancées telles que la limite de mémoire ou du code SQL personnalisé. [Pour plus d'informations, reportez-vous à la section Options de puissance sur p. 121.](#)

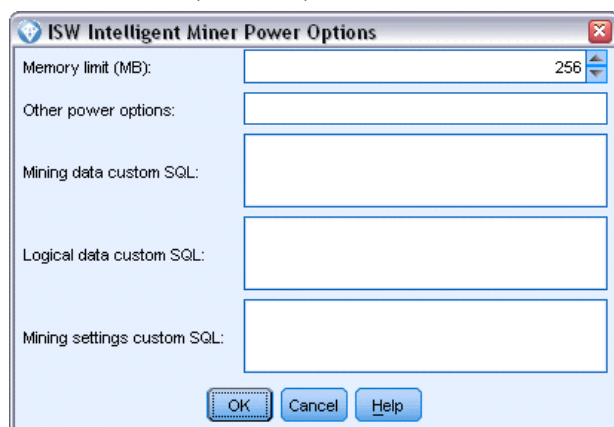
L'onglet Serveur d'un noeud généré inclut une option qui permet de vérifier la cohérence en stockant une chaîne de clé de modèle généré identique dans les modèles SPSS Modeler et DB2. [Pour plus d'informations, reportez-vous à la section Nugget de modèle ISW - Onglet Serveur sur p. 155.](#)

### Options de puissance

L'onglet Serveur de tous les algorithmes comporte une case à cocher permettant d'activer les options de puissance de la modélisation ISW. Lorsque vous cliquez sur le bouton Options de puissance, vous voyez apparaître la boîte de dialogue Options de puissance d'ISW, comportant les options suivantes :

- Limite de mémoire.
- Autres options de puissance.
- Code SQL personnalisé des données d'exploration.
- Code SQL personnalisé de données logiques.
- Code SQL personnalisé des paramètres d'exploration.

Figure 5-8  
Paramètres des options de puissance d'ISW



**Limite de mémoire** Limite la consommation de mémoire d'un algorithme de création de modèle. L'option de puissance standard définit une limite sur le nombre de valeurs discrètes dans les données catégorielles.

**Autres options de puissance.** Permet d'indiquer des options de puissance arbitraires sous forme de lignes de commande pour certains modèles ou solutions. Ces éléments spécifiques varient en fonction de la mise en oeuvre ou de la solution. Vous pouvez développer manuellement le SQL généré par IBM® SPSS® Modeler pour définir une tâche de création de modèle.

**Code SQL personnalisé des données d'exploration.** Vous pouvez ajouter des appels de méthode pour modifier l'objet `DM_MiningData`. Par exemple, si vous entrez le SQL suivant, un filtre basé sur un champ appelé *Partition* est ajouté aux données utilisées pour la création du modèle :

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

**Code SQL personnalisé de données logiques.** Vous pouvez ajouter des appels de méthode pour modifier l'objet `DM_LogicalDataSpec`. Par exemple, le SQL suivant supprime un champ de l'ensemble de champs utilisé pour la création du modèle :

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

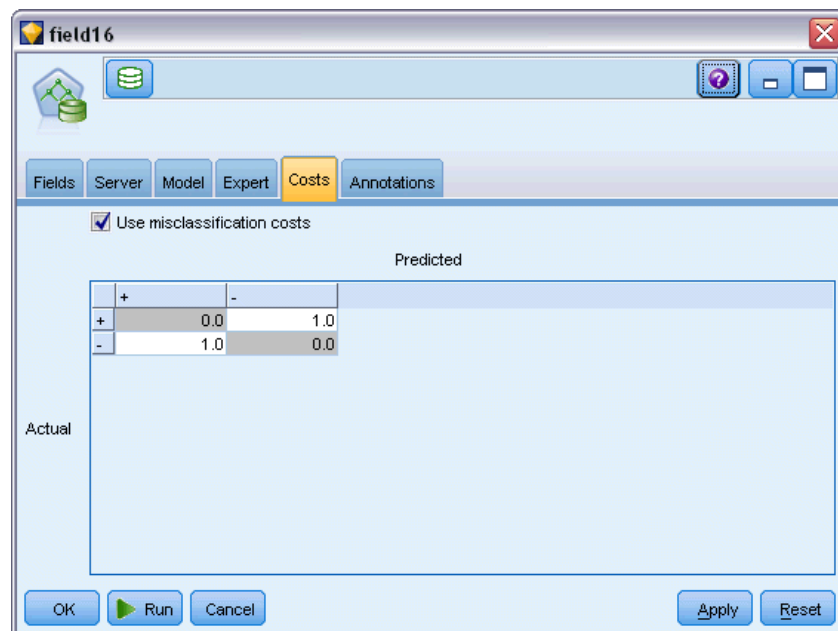
**Code SQL personnalisé des paramètres d'exploration.** Vous pouvez ajouter des appels de méthode pour modifier l'objet `DM_ClasSettings/DM_RuleSettings/DM_ClusSettings/DM_RegSettings`. Par exemple, si vous entrez le SQL suivant, IBM InfoSphere Warehouse Data Mining active le champ *Partition* (ce qui signifie qu'il est toujours inclus dans le modèle résultant) :

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

### Options des coûts d'ISW

Dans l'onglet Coûts, vous pouvez ajuster les coûts de mauvaise réaffectation, ce qui vous permet d'indiquer l'importance relative de différentes erreurs de prévision.

Figure 5-9  
Onglets Coûts d'ISW



Selon le contexte, certains types d'erreur peuvent se révéler plus coûteux que d'autres. Par exemple, il peut être plus coûteux de classer un candidat au crédit à haut risque dans la catégorie à faible risque (un type d'erreur) que de classer un candidat à faible risque dans la catégorie à haut risque (un autre type d'erreur). L'option des coûts de mauvaise réaffectation vous permet de spécifier l'importance relative de différentes erreurs de prévision.

Les coûts d'une mauvaise réaffectation sont des pondérations appliquées à des revenus définis. Elles sont prises en compte dans le modèle et peuvent modifier la prévision (ce qui permet d'éviter des erreurs qui pourraient coûter cher).

A l'exception des modèles C5.0, les coûts d'une mauvaise réaffectation ne s'appliquent lorsque vous évaluez un modèle et ne sont pas pris en compte lors du classement ou de la comparaison de modèles par le biais d'un noeud Classificateur automatique, d'un graphique Evaluation ou d'un noeud Analyse. Il se peut qu'un modèle comprenant des coûts ne produise pas moins d'erreurs qu'un modèle n'en comprenant pas et ne classe pas de façon maximale en termes de précision générale. En revanche, il est probable, qu'en pratique, ses performances soient meilleures du fait qu'il dispose de biais intégrés rendant les erreurs *moins coûteuses*.

La matrice de mauvaise réaffectation des coûts affiche le coût de chaque combinaison possible de catégories prédites et de catégories réelles. Par défaut, tous les coûts de mauvaise réaffectation sont paramétrés sur 1. Pour entrer des valeurs de coût personnalisées, sélectionnez Utiliser les coûts de mauvaise réaffectation et entrez vos valeurs personnalisées dans la matrice des coûts.

Pour modifier un coût dû à une mauvaise réaffectation, sélectionnez la cellule correspondant à la combinaison voulue de valeurs prédites et de valeurs réelles, supprimez le contenu de la cellule et entrez le coût à appliquer à la cellule. Les coûts ne sont pas automatiquement symétriques. Ainsi, si vous définissez le coût d'une mauvaise affectation de *A* en tant que *B* sur 2, le coût d'une mauvaise affectation de *B* en tant que *A* sera toujours défini sur la valeur par défaut 1, à moins que vous ne modifiez cette valeur de manière explicite.

## **Arbre décision ISW**

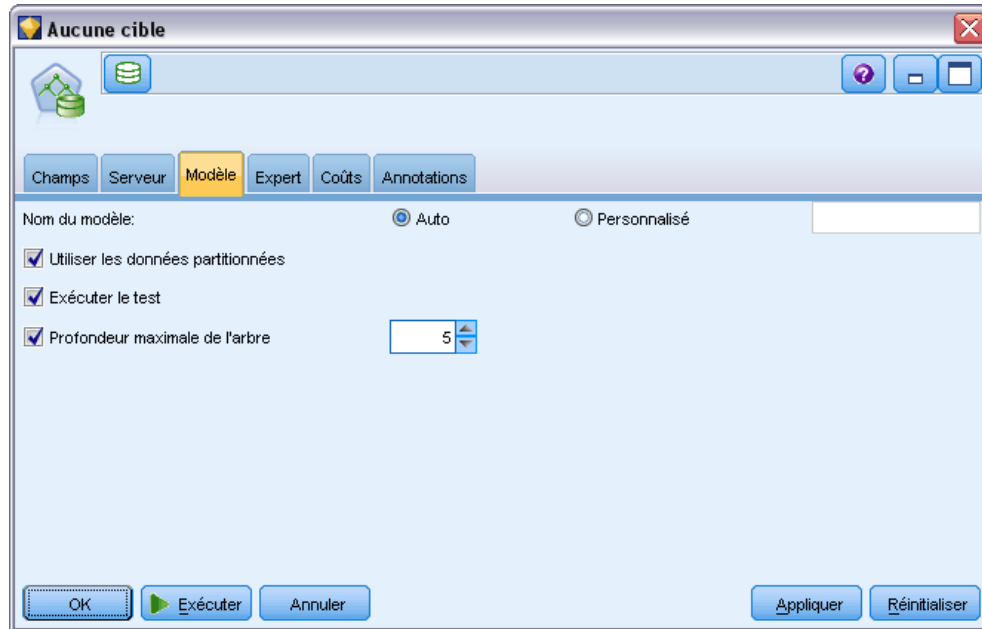
Les modèles d'arbre décision vous permettent de développer des systèmes de classification prévoyant ou classant les observations futures à partir d'un ensemble de règles de décision. Si vos données sont divisées en classes qui vous intéressent (par exemple, rapport prêts à haut risque, prêts à faible risque, abonnés/non-abonnés, votants/abstentionnistes, ou types de bactérie), vous pouvez les utiliser pour construire des règles qui permettront de classer les observations anciennes ou nouvelles avec une précision maximale. Par exemple, vous pouvez construire un arbre qui classe le risque de crédit ou l'intention d'achat en fonction de l'âge et d'autres facteurs.

L'algorithme Arbre décision ISW crée des arbres de classification sur des données d'entrée catégorielles. L'arbre décision obtenu est binaire. Vous pouvez appliquer divers paramètres, comme des coûts de mauvaise réaffectation, lors de la création du modèle.

L'outil ISW Visualizer est la seule méthode permettant de parcourir les modèles IBM InfoSphere Warehouse Data Mining.

## Arbre décision ISW - Options du modèle

Figure 5-10  
Noeud Arbre décision ISW, onglet Modèle



**Nom du modèle.** Vous pouvez générer le nom du modèle automatiquement sur la base du champ cible ou ID (ou du type de modèle si aucun de ces champs n'est spécifié) ou spécifier un nom personnalisé.

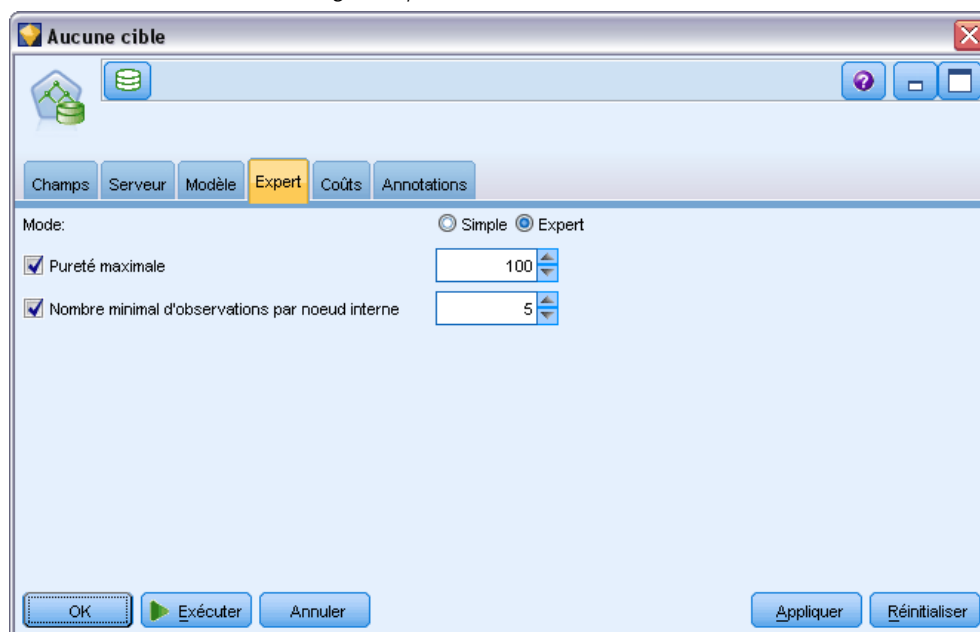
**Utiliser les données partitionnées.** Si vous définissez un champ de partition, sélectionnez Utiliser les données partitionnées.

**Exécuter le test.** Vous pouvez réaliser un test. Un test IBM InfoSphere Warehouse Data Mining est alors exécuté une fois le modèle créé sur la partition d'apprentissage. Cela engendre un passage sur la partition test, permettant d'établir des informations sur la qualité du modèle, des graphiques Lift, etc.

**Profondeur maximale de l'arbre.** Vous pouvez indiquer la profondeur d'arbre maximale. Cette valeur limite la profondeur de l'arbre au nombre de niveaux indiqué. Si cette option n'est pas sélectionnée, aucune limite n'est appliquée. Pour éviter la création de modèles trop complexes, une valeur supérieure à 5 n'est généralement pas recommandée.

## Arbre de décision ISW - Options expert

Figure 5-11  
Noeud Arbre décision ISW, onglet Expert



**Pureté maximale.** Cette option définit la pureté maximale des noeuds internes. Si l'un des noeuds enfant résultant de la division d'un noeud dépasse la mesure de pureté indiquée (par exemple, plus de 90 % des observations appartiennent à une catégorie spécifique), le noeud n'est pas divisé.

**Nombre minimal d'observations par noeud interne.** Si la division d'un noeud engendre un enfant comportant moins d'observations que le nombre minimal indiqué, le noeud n'est pas divisé.

## Association ISW

Vous pouvez utiliser le noeud d'association de ISW pour rechercher des règles d'association parmi les éléments présents dans un ensemble de groupes. Les règles d'association associent une conclusion particulière (par exemple, l'achat d'un produit particulier) à un ensemble de conditions (par exemple, l'achat de plusieurs autres produits).

Vous pouvez choisir d'inclure ou d'exclure des règles d'association du modèle en spécifiant les **contraintes**. Si vous choisissez d'inclure un champ d'entrée particulier, les règles d'association qui contiennent au moins un des éléments spécifiés sont incluses dans le modèle. Si vous excluez un champ d'entrée, les règles d'association qui contiennent un des éléments spécifiés ne sont pas prises en compte dans les résultats.

Les algorithmes d'association et de séquence de ISW peuvent utiliser les **taxonomies**. Les taxonomies mappent chaque valeur à un concept de niveau supérieur. Par exemple, stylos et crayons peuvent être mappés à la catégorie Papeterie.

Les règles d'association ont une seule conséquence (la conclusion) et plusieurs antécédents (l'ensemble des conditions). Voir l'exemple comme suit :

[Bread, Jam] • [Butter]

[Bread, Jam]  
• [Margarine]

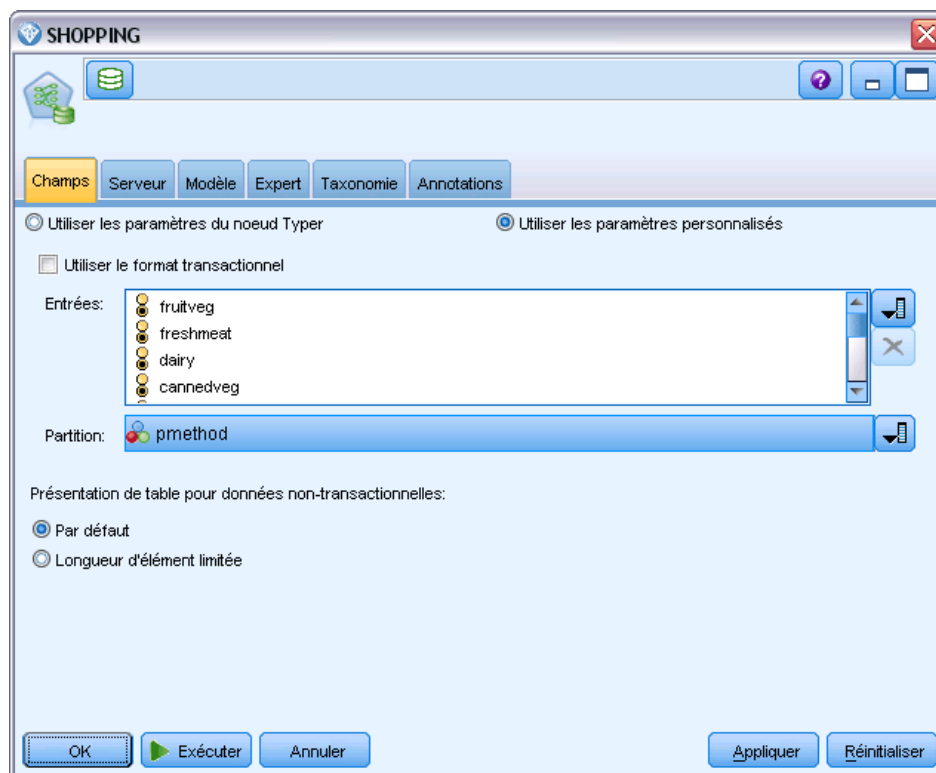
Ici, Bread et Jam sont les antécédents (également appelés **corps de la règle**) et Butter ou Margarine sont tous les deux des exemples de conséquence (également appelée **en-tête de règle**). La première règle indique qu'une personne qui a acheté du pain et de la confiture a également acheté du beurre en même temps. La deuxième règle identifie un client qui, lorsqu'il a acheté cette même combinaison (pain et confiture) a également acheté de la margarine lors d'une même visite dans le magasin.

L'outil Visualizer est la seule méthode permettant de parcourir les modèles IBM InfoSphere Warehouse Data Mining.

## Association ISW - Options de champs

L'onglet Champs vous permet de spécifier les champs à utiliser lors de la création du modèle.

Figure 5-12  
Noeud Association ISW - Onglet Champs





Avant de construire un modèle, vous devez indiquer les champs à utiliser en tant que cibles et en tant qu'entrées. A quelques exceptions près, tous les noeuds de modélisation utilisent les informations de champ d'un nœud Typer en amont. En dehors du paramètre par défaut qui définit l'utilisation du nœud Typer pour sélectionner les champs d'entrée et les champs cible, le seul autre paramètre que vous pouvez modifier sur cet onglet est la présentation de table des données non-transactionnelles.

**Utiliser les paramètres du nœud Typer.** Cette option indique au système d'utiliser les informations de champ provenant d'un nœud Typer en amont. Il s'agit de la valeur par défaut.

**Utiliser les paramètres personnalisés.** Cette option indique au système d'utiliser les informations de champ saisies ici au lieu des informations provenant d'un nœud Typer en amont. Une fois cette option sélectionnée, renseignez les champs ci-dessous.

**Utilisation du format transactionnel.** Sélectionnez cette case si les données source sont au **format transactionnel**. Les enregistrements dans ce format ont deux champs, un pour l'ID et l'autre pour le contenu. Chaque enregistrement représente une seule transaction ou un seul élément et les éléments associés sont liés en ayant le même ID. Désélectionnez cette case si les données sont au **format tabulaire**, dans lequel les éléments sont représentés par des éléments booléens distincts, où chaque champ booléen représente la présence ou l'absence d'un élément spécifique et chaque enregistrement représente un ensemble d'éléments associés complet. [Pour plus d'informations, reportez-vous à la section Données tabulaires et données transactionnelles dans le chapitre 12 dans Noeuds de modélisation de IBM SPSS Modeler 15.](#)

- **ID.** Pour des données transactionnelles, sélectionnez un champ ID dans la liste. Vous pouvez utiliser un champ numérique ou symbolique en tant que champ ID. Chaque valeur unique de ce champ doit indiquer une unité d'analyse spécifique. Par exemple, dans une analyse de paniers du marché, chaque ID peut représenter un client. Dans une analyse de l'utilisation du Web, chaque ID peut représenter un ordinateur (adresse IP) ou un utilisateur (ID de connexion).
- **Contenu.** Spécifiez le champ de contenu ou les champs du modèle. Ces champs contiennent les éléments d'intérêt concernant la modélisation des associations. Vous pouvez spécifier un champ nominal unique si les données sont au format transactionnel.

**Utilisation du format tabulaire.** Désélectionnez la case Utiliser le format transactionnel si les données source sont au format tabulaire.

- **Entrées.** Sélectionnez le ou les champs d'entrée. Cela revient à définir le rôle du champ sur la valeur *Entrée* dans un nœud Typer.
- **Partition.** Ce champ permet d'indiquer un champ utilisé pour partitionner les données en échantillons distincts pour les étapes d'apprentissage, de test et de validation de la création d'un modèle. L'utilisation d'un échantillon pour la génération du modèle et d'un échantillon distinct pour le tester vous permet d'avoir une bonne indication de la manière dont le modèle peut se généraliser à des ensembles de données plus importants, similaires aux données actuelles. Si plusieurs champs de partition sont définis via des noeuds Typer ou Partitionner, vous devez en sélectionner un seul dans l'onglet Champs de chaque nœud de modélisation ayant recours au partitionnement. (Dans le cas d'une seule partition, cette partition est automatiquement utilisée lorsque la fonction de partition est activée.) [Pour plus d'informations, reportez-vous à la section Nœud Partitionner dans le chapitre 4 dans Noeuds source, exécution et de sortie de IBM SPSS Modeler 15.](#) Notez également que, pour appliquer la partition sélectionnée à l'analyse, vous devez activer l'option de partitionnement dans

l'onglet Options de modèle du noeud. (Désélectionnez cette option pour pouvoir désactiver la partition sans modifier les paramètres du champ.)

**Présentation de table pour les données non-transactionnelles.** Si les données sont tabulaires, vous pouvez choisir une présentation de table standard (par défaut) ou une présentation de longueur limitée.

Dans la présentation par défaut, le nombre de colonnes est déterminé par le nombre total des éléments associés.

Table 5-2  
*Présentation de table par défaut*

ID de groupe	Compte chèque	Compte d'épargne	Carte de crédit	Prêt	Compte de dépôt
Smith	Y	Y	Y	-	-
Jackson	Y	-	Y	Y	Y
Douglas	Y	-	-	-	Y

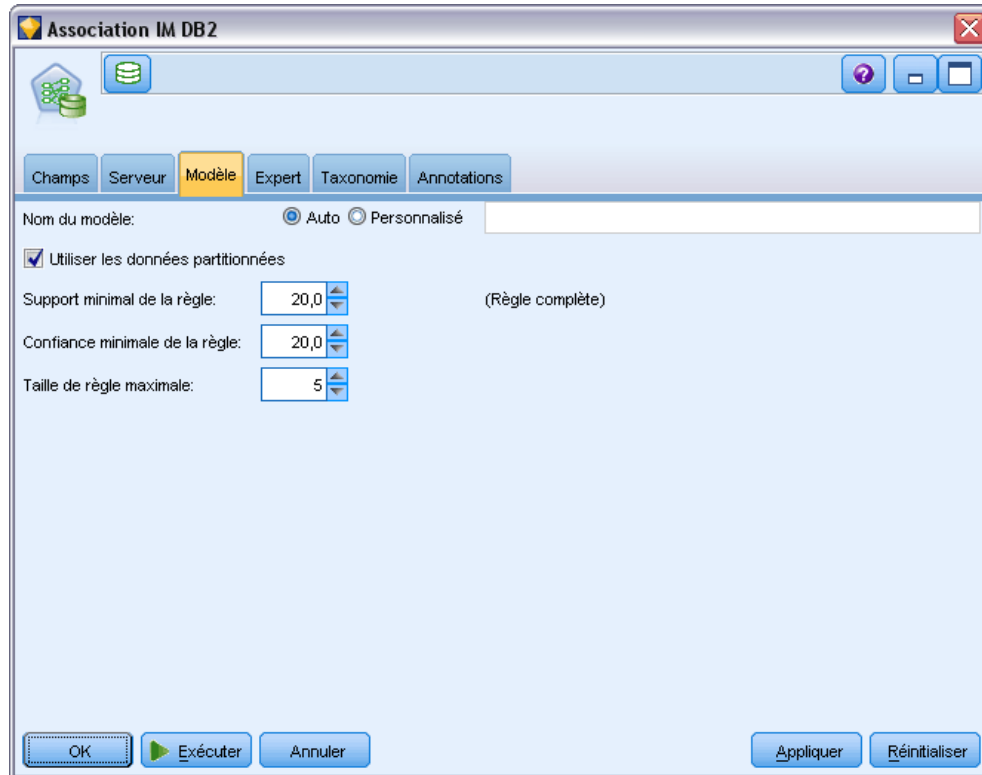
Dans la présentation de longueur limitée, le nombre de colonnes est déterminé par le plus grand nombre d'éléments associés à une ligne.

Table 5-3  
*Présentation de table de longueur limitée*

ID de groupe	Élément 1	Élément 2	Élément 3	Élément 4
Smith	compte chèque	compte d'épargne	carte de crédit	-
Jackson	compte chèque	carte de crédit	prêt	compte de dépôt
Douglas	compte chèque	compte de dépôt	-	-

## Association ISW - Options de modèle

Figure 5-13  
Noeud Association ISW - Onglet Modèle



**Nom du modèle.** Vous pouvez générer le nom du modèle automatiquement sur la base du champ cible ou ID (ou du type de modèle si aucun de ces champs n'est spécifié) ou spécifier un nom personnalisé.

**Utiliser les données partitionnées.** Si un champ de partition est défini, cette option assure que seules les données provenant de la partition d'apprentissage sont utilisées pour générer le modèle. [Pour plus d'informations, reportez-vous à la section Noeud Partitionner dans le chapitre 4 dans Noeuds source, exécution et de sortie de IBM SPSS Modeler 15.](#)

**Prise en charge minimale de la règle (%).** Niveau de prise en charge minimale des règles d'association ou de séquence. Seules les règles atteignant au moins ce niveau de prise en charge sont incluses dans le modèle. La valeur est calculée comme  $A/B*100$  où A est le nombre de groupes contenant tous les éléments qui apparaissent dans la règle et B est le nombre total de tous les groupes pris en compte. Si vous souhaitez extraire des associations ou séquences plus standard, augmentez la valeur de ce paramètre.

**Confiance minimale de la règle (%).** Niveau de confiance minimale des règles d'association ou de séquence. Seules les règles atteignant au moins ce niveau de confiance sont incluses dans le modèle. La valeur est calculée comme  $m/n*100$ , où m est le nombre de groupes contenant l'en-tête de règle (conséquence) et le corps de règle liés (antécédent) et n est le nombre de groupes contenant le corps de la règle. Si vous obtenez des associations ou des séquences inintéressantes

ou trop nombreuses, augmentez la valeur de ce paramètre. En revanche, réduisez la valeur si le nombre d'associations ou de séquences n'est pas assez important.

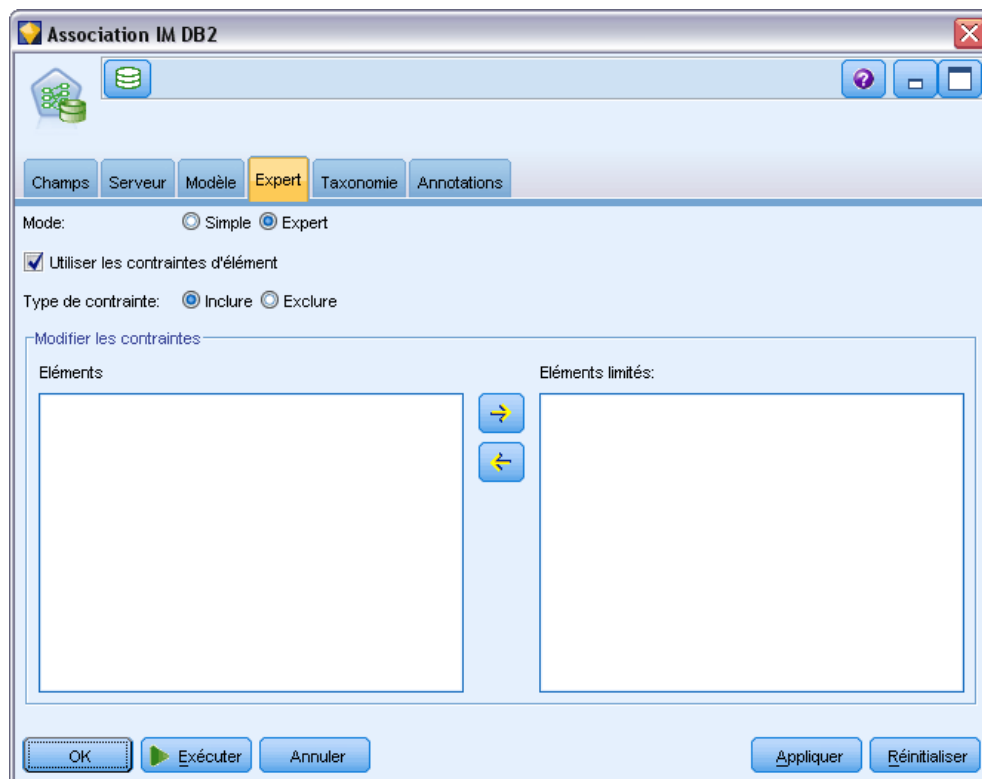
**Taille de règle maximale.** Nombre maximal d'éléments autorisés dans une règle, y compris l'élément de conséquence. Si les associations ou les séquences qui vous intéressent sont relativement courtes, diminuez la valeur de ce paramètre afin d'accélérer la génération de l'ensemble.

*Remarque* : Seuls les noeuds dont le format d'entrée est transactionnel sont évalués, les formats tabulaires des valeurs vraies (données tabulaires) restant non affinés.

## Association ISW - Options expert

Dans l'onglet Expert du noeud Association, vous pouvez spécifier les règles d'association à inclure dans les résultats, ou à exclure des résultats. Si vous décidez d'inclure des éléments spécifiés, les règles qui contiennent au moins un des éléments spécifiés sont incluses dans le modèle. Si vous décidez d'exclure des éléments spécifiés, les règles qui contiennent un ou plusieurs éléments spécifiés ne sont pas prises en compte dans les résultats.

Figure 5-14  
Noeud Association ISW - Onglet Expert



Lorsque l'option Utiliser les contraintes d'élément est sélectionnée, tous les éléments ajoutés à la liste des contraintes sont inclus ou exclus des résultats, selon le type de contrainte défini.

**Type de contrainte.** Choisissez si vous voulez inclure ou exclure des résultats ces règles d'association qui contiennent les éléments spécifiés.

**Modifier les contraintes.** Pour ajouter un élément à la liste des éléments limités, sélectionnez-le dans la liste Eléments et cliquez sur la flèche droite.

### **Options de taxonomie d'ISW**

Les algorithmes d'association et de séquence de ISW peuvent utiliser les **taxonomies**. Les taxonomies mappent chaque valeur à un concept de niveau supérieur. Par exemple, stylos et crayons peuvent être mappés à la catégorie Papeterie.

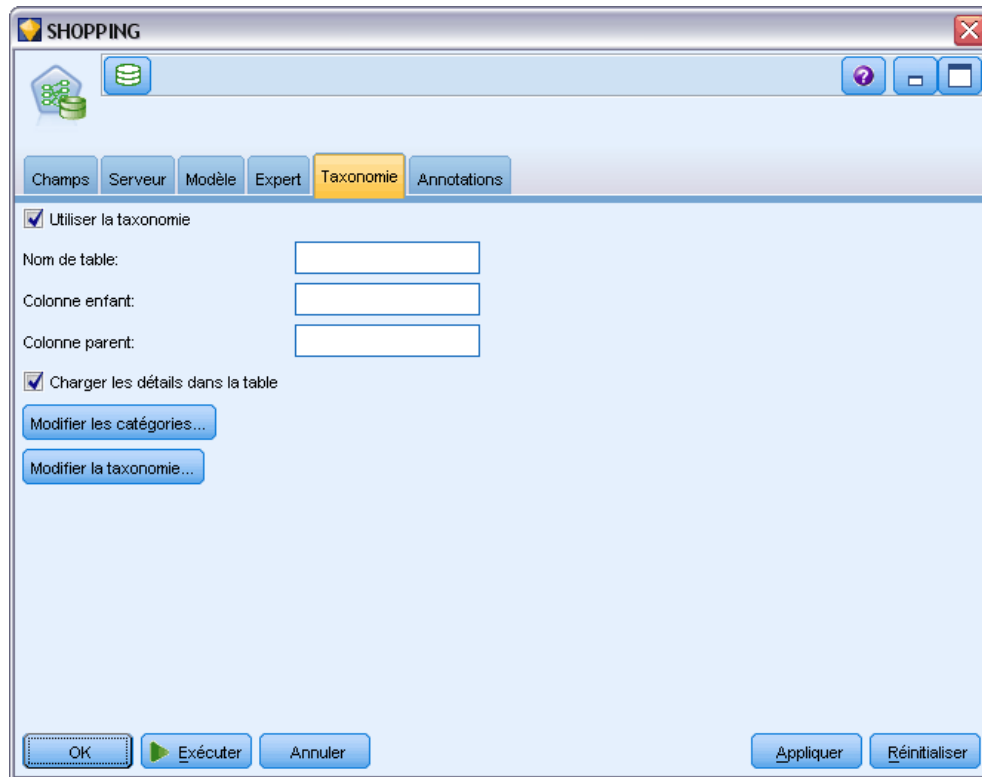
Dans l'onglet Taxonomie, vous pouvez définir des mappages de catégorie afin d'exprimer des taxonomies entre les données. Supposons, par exemple, qu'une taxonomie crée deux catégories (Staple et Luxury), puis affecte des éléments principaux à chaque catégorie. Par exemple, wine est affecté à Luxury et bread est affecté à Staple. La taxonomie a une structure parent-enfant comme décrit ci-dessous :

Enfant	Parent
vin	Produit de luxe
pain	Produit de base

Cette taxonomie vous permet de créer un modèle d'association ou de séquence qui inclut des règles impliquant les catégories, ainsi que les éléments principaux.

*Remarque* : Pour activer les options de cet onglet, les données source doivent être au format transactionnel et vous devez sélectionner Utiliser le format transactionnel dans l'onglet Champs puis sélectionner Utiliser la taxonomie sur cet onglet. [Pour plus d'informations, reportez-vous à la section Données tabulaires et données transactionnelles dans le chapitre 12 dans Noeuds de modélisation de IBM SPSS Modeler 15.](#)

Figure 5-15  
Noeud Association ISW - Onglet Taxonomie



**Nom de la table.** Cette option indique le nom de la table DB2 dans laquelle stocker les détails de la taxonomie.

**Colonne enfant.** Cette option indique le nom de la colonne enfant de la table de taxonomie. La colonne enfant comporte les noms des éléments ou des catégories.

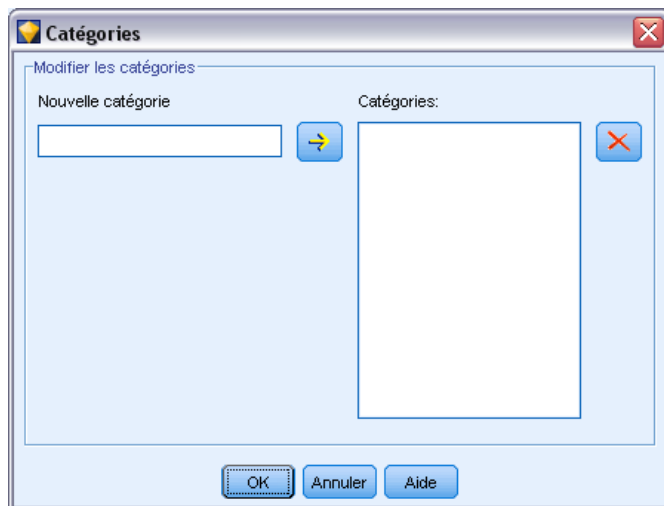
**Colonne parent.** Cette option indique le nom de la colonne parent de la table de taxonomie. La colonne parent contient les noms des catégories.

**Charger les détails dans la table.** Sélectionnez cette option si les informations sur la taxonomie stockées dans IBM® SPSS® Modeler doivent être envoyées vers la table de taxonomie lors de la création du modèle. La table de taxonomie est supprimée si elle existe déjà. Les informations sur la taxonomie sont stockées avec le noeud de création du modèle et éditées via les boutons Modifier les catégories et Modifier la taxonomie.

### ***Editeur de catégorie***

La boîte de dialogue Modifier les catégories permet d'ajouter des catégories à une liste triée et d'en supprimer.

Figure 5-16  
Editeur de catégories de taxonomie



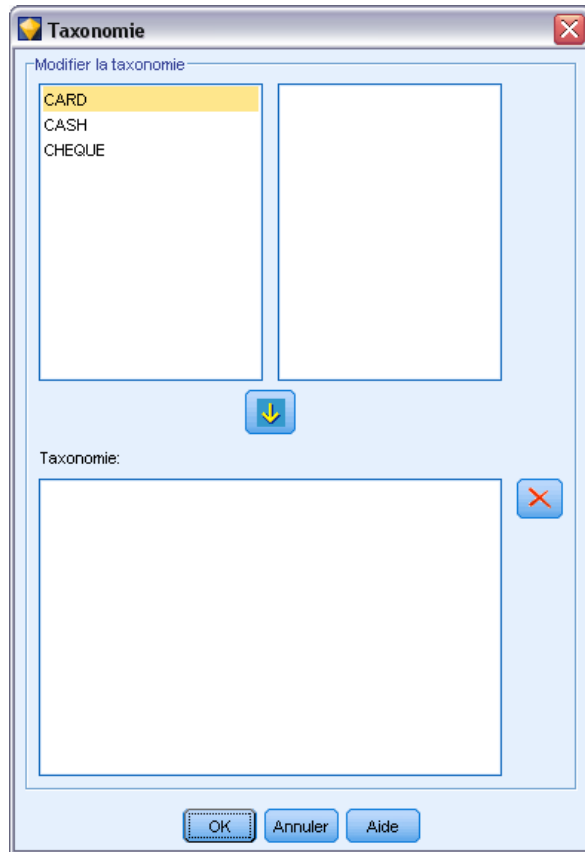
Pour ajouter une catégorie, saisissez son nom dans le champ Nouvelle catégorie puis cliquez sur la flèche pour le déplacer dans la liste Catégories.

Pour supprimer une catégorie, sélectionnez-la dans la liste Catégories et cliquez sur le bouton adjacent Supprimer.

### **Editeur de taxonomie**

La boîte de dialogue Modifier la taxonomie permet de combiner le jeu d'éléments principaux défini dans les données et l'ensemble de catégories pour la création d'une taxonomie. Pour ajouter des entrées à la taxonomie, sélectionnez des éléments ou catégories dans la liste de gauche et des catégories dans la liste de droite, puis cliquez sur le bouton fléché. Lorsque des ajouts dans la taxonomie entraînent un conflit (par exemple lorsque vous indiquez `cat1 -> cat2` et le contraire, `cat2 -> cat1`), ils ne sont pas réalisés.

Figure 5-17  
Editeur de taxonomie



## Séquence ISW

Le noeud Séquence recherche des motifs dans des données séquentielles ou des données liées au temps, au format `bread -> cheese`. Les composants d'une séquence sont des **jeux d'éléments** constituant une transaction unique. Supposons, par exemple, qu'une personne aille au supermarché et achète du pain et du lait, puis retourne quelques jours plus tard au supermarché pour acheter du fromage. Les achats de cette personne peuvent alors être représentés par deux jeux d'éléments. Le premier jeu contient le pain et le lait, le second jeu contient le fromage. Une **séquence** est une liste de jeux d'éléments ayant tendance à survenir dans un ordre prévisible. Le noeud Séquence détecte les séquences les plus fréquentes et crée un noeud de modèle généré pouvant être utilisé pour établir des prévisions.

Vous pouvez utiliser la fonction de Data mining Règles de séquence dans différents secteurs de marché. Par exemple, dans le secteur de la vente au détail, vous pouvez trouver des séries types d'achats. Ces séries montrent les différentes combinaisons clients/produits/période d'achat. Grâce à ces informations, vous pouvez identifier les clients potentiels susceptibles d'acheter un produit particulier pour la première fois. En outre, vous pouvez offrir des produits aux clients potentiels en temps voulu.

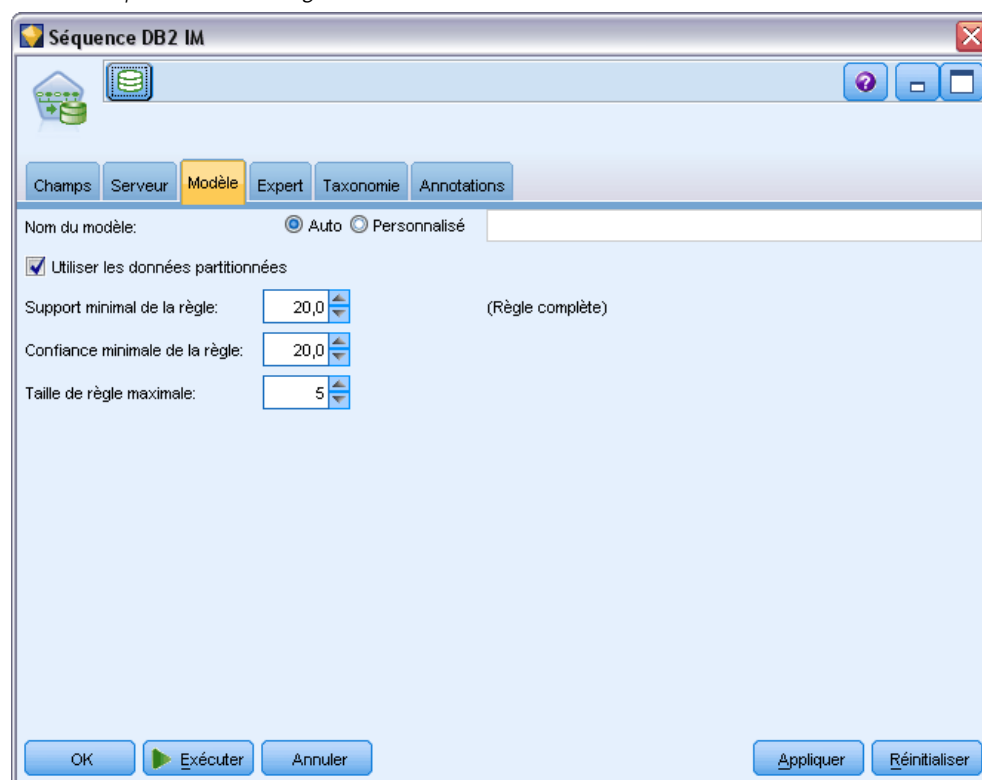


Une séquence constitue un ensemble ordonné de jeux d'éléments. Les séquences contiennent les niveaux de regroupement suivants :

- Les événements qui ont lieu simultanément forment une transaction unique ou un jeu d'éléments.
- Chaque élément ou chaque jeu d'éléments appartient à un groupe de transactions. Par exemple, un article acheté appartient à un client, un clic sur une page spécifique appartient à un utilisateur Internet ou un composant appartient à un véhicule produit. Plusieurs jeux d'éléments qui ont lieu à différents moments et qui appartiennent au même groupe de transactions forment une séquence.

## Séquence ISW - Options de modèle

Figure 5-18  
Noeud Séquence ISW - Onglet Modèle



**Nom du modèle.** Vous pouvez générer le nom du modèle automatiquement sur la base du champ cible ou ID (ou du type de modèle si aucun de ces champs n'est spécifié) ou spécifier un nom personnalisé.

**Utiliser les données partitionnées.** Si un champ de partition est défini, cette option assure que seules les données provenant de la partition d'apprentissage sont utilisées pour générer le modèle. [Pour plus d'informations, reportez-vous à la section Noeud Partitionner dans le chapitre 4 dans Noeuds source, exécution et de sortie de IBM SPSS Modeler 15.](#)

**Prise en charge minimale de la règle (%).** Niveau de prise en charge minimale des règles d'association ou de séquence. Seules les règles atteignant au moins ce niveau de prise en charge sont incluses dans le modèle. La valeur est calculée comme  $A/B*100$  où A est le nombre de groupes contenant tous les éléments qui apparaissent dans la règle et B est le nombre total de tous les groupes pris en compte. Si vous souhaitez extraire des associations ou séquences plus standard, augmentez la valeur de ce paramètre.

**Confiance minimale de la règle (%).** Niveau de confiance minimale des règles d'association ou de séquence. Seules les règles atteignant au moins ce niveau de confiance sont incluses dans le modèle. La valeur est calculée comme  $m/n*100$ , où  $m$  est le nombre de groupes contenant l'en-tête de règle (conséquence) et le corps de règle liés (antécédent) et  $n$  est le nombre de groupes contenant le corps de la règle. Si vous obtenez des associations ou des séquences inintéressantes ou trop nombreuses, augmentez la valeur de ce paramètre. En revanche, réduisez la valeur si le nombre d'associations ou de séquences n'est pas assez important.

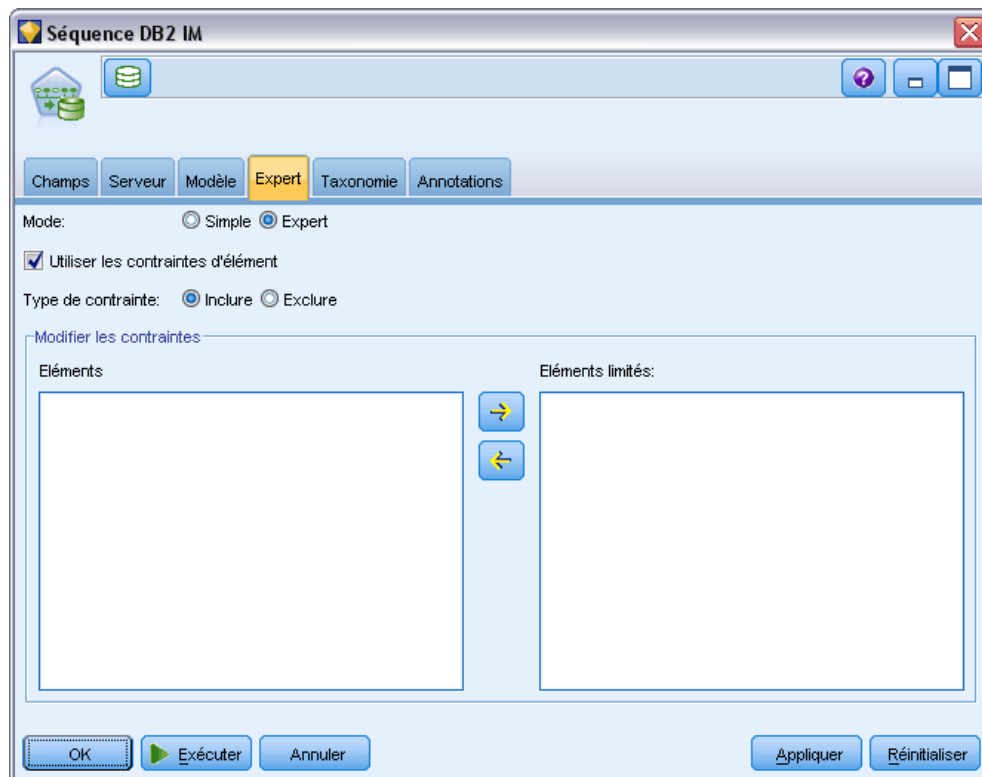
**Taille de règle maximale.** Nombre maximal d'éléments autorisés dans une règle, y compris l'élément de conséquence. Si les associations ou les séquences qui vous intéressent sont relativement courtes, diminuez la valeur de ce paramètre afin d'accélérer la génération de l'ensemble.

*Remarque :* Seuls les noeuds dont le format d'entrée est transactionnel sont évalués, les formats tabulaires des valeurs vraies (données tabulaires) restant non affinés.

### ***Séquence ISW - Options expert***

Vous pouvez spécifier les règles de séquence à inclure dans les résultats, ou à exclure des résultats. Si vous décidez d'inclure des éléments spécifiés, les règles qui contiennent au moins un des éléments spécifiés sont incluses dans le modèle. Si vous décidez d'exclure des éléments spécifiés, les règles qui contiennent un ou plusieurs éléments spécifiés ne sont pas prises en compte dans les résultats.

Figure 5-19  
Noeud Séquence ISW - Onglet Expert



Lorsque l'option Utiliser les contraintes d'élément est sélectionnée, tous les éléments ajoutés à la liste des contraintes sont inclus ou exclus des résultats, selon le type de contrainte défini.

**Type de contrainte.** Choisissez si vous voulez inclure ou exclure des résultats ces règles d'association qui contiennent les éléments spécifiés.

**Modifier les contraintes.** Pour ajouter un élément à la liste des éléments limités, sélectionnez-le dans la liste Eléments et cliquez sur la flèche droite.

## Régression ISW

Le noeud Régression ISW prend en charge les algorithmes de régression suivants :

- Transformation (par défaut)
- Linéaire
- Polynomial
- RBF

### ***Régression de transformation***

L'algorithme de régression de transformation ISW crée des modèles qui sont des arbres décision comportant des équations de régression sur leurs feuilles. Veuillez noter qu'IBM Visualizer n'affiche pas la structure de ces modèles.

Le navigateur IBM® SPSS® Modeler affiche les paramètres et les annotations. Cependant, vous ne pouvez pas parcourir la structure du modèle. Il existe relativement peu de paramètres de création pouvant être configurés par l'utilisateur.

### ***Régression linéaire***

L'algorithme de régression linéaire ISW suppose une relation linéaire entre les champs explicatifs et le champ cible. Il produit des modèles qui représentent des équations. La valeur prédite est censée être différente de la valeur observée car une équation de régression est une approximation du champ cible. La différence est appelée résidu.

La modélisation IBM InfoSphere Warehouse Data Mining reconnaît les champs qui ne présentent pas de valeur explicative. Pour déterminer si un champ présente une valeur explicative, l'algorithme de régression linéaire exécute des tests statistiques en plus de la sélection automatique de variables. Si vous connaissez les champs qui ne présentent pas ce type de valeur explicative, vous pouvez sélectionner automatiquement un sous-ensemble de champs explicatifs pour des durées d'exécution plus courtes.

L'algorithme de régression linéaire fournit les méthodes suivantes pour sélectionner automatiquement des sous-ensembles de champs explicatifs :

**Régression Pas à pas.** Pour la régression Pas à pas, vous devez spécifier un niveau de signification minimal. Seuls les champs qui présentent un niveau de signification supérieur à la valeur spécifiée sont utilisés par l'algorithme de régression linéaire.

**Régression R carré.** La méthode de régression R carré identifie un modèle optimal en optimisant une mesure de la qualité du modèle. L'une des mesures de qualité suivantes est utilisée :

- Le coefficient de corrélation de Pearson au carré
- Le coefficient de corrélation de Pearson au carré ajusté.

Par défaut, l'algorithme de régression linéaire sélectionne automatiquement des sous-ensembles de champs explicatifs en utilisant le coefficient de corrélation de Pearson au carré ajusté pour optimiser la qualité du modèle.

### ***Régression polynomiale***

L'algorithme de régression polynomiale ISW suppose une relation polynomiale. Un modèle de régression polynomiale est une équation qui se compose des éléments suivants :

- Le degré maximal de régression polynomiale
- Une approximation du champ cible
- Les champs explicatifs.

### Régression RBF

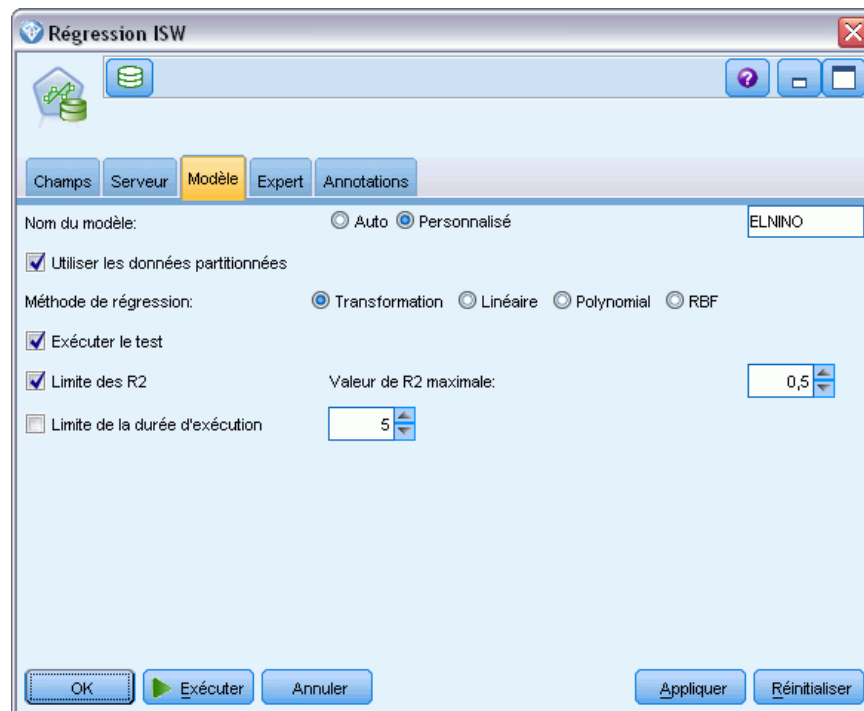
L'algorithme de régression RBF ISW suppose une relation entre les champs explicatifs et le champ cible. Cette relation peut être exprimée par une combinaison linéaire des fonctions gaussiennes. Les fonctions gaussiennes sont des fonctions radiales de base.

## Régression ISW - Options de modèle

Dans l'onglet Modèle du noeud Régression ISW, vous pouvez spécifier le type d'algorithme de régression à utiliser ainsi que :

- Si vous voulez utiliser des données partitionnées
- Si vous voulez réaliser un test
- Une limite pour la valeur  $R^2$
- Une limite pour le temps d'exécution

Figure 5-20  
Noeud Régression ISW - Onglet Modèle



**Utiliser les données partitionnées.** Si un champ de partition est défini, cette option assure que seules les données provenant de la partition d'apprentissage sont utilisées pour générer le modèle. [Pour plus d'informations, reportez-vous à la section Noeud Partitionner dans le chapitre 4 dans Noeuds source, exécution et de sortie de IBM SPSS Modeler 15.](#)

**Méthode de régression.** Choisissez le type de régression à effectuer. [Pour plus d'informations, reportez-vous à la section Régression ISW sur p. 137.](#)

**Exécuter le test.** Vous pouvez réaliser un test. Un test InfoSphere Warehouse Data Mining est alors exécuté une fois le modèle créé sur la partition d'apprentissage. Cela engendre un passage sur la partition test, permettant d'établir des informations sur la qualité du modèle, des graphiques Lift, etc.

**Limite des R2.** Cette option indique l'erreur systématique maximale tolérée (le coefficient de corrélation de Pearson au carré,  $R^2$ ). Ce coefficient mesure la corrélation entre l'erreur de prédiction dans les données de vérification et les valeurs cibles réelles. Il contient une valeur entre 0 (aucune corrélation) et 1 (corrélation positive ou négative parfaite). La valeur définie ici indique la limite supérieure acceptable de l'erreur systématique du modèle.

**Limite de la durée d'exécution.** Indiquez, en minutes, le temps d'exécution maximal souhaité.

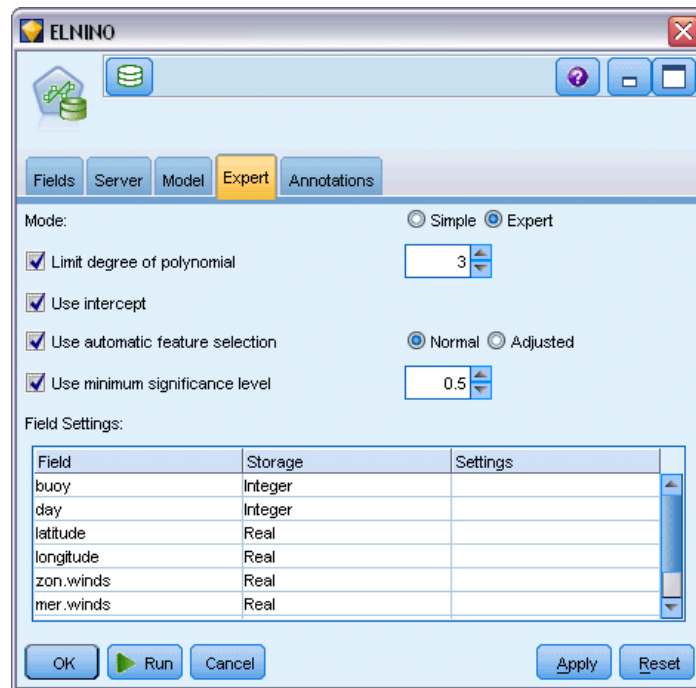
## Régression ISW - Options expert

Dans l'onglet Expert du noeud Régression ISW, vous pouvez spécifier un nombre d'options avancées pour la régression linéaire, polynomiale ou RBF.

### Options avancées de la régression linéaire ou polynomiale

Figure 5-21

Noeud Régression ISW, Onglet Expert pour la régression linéaire ou polynomiale



**Degré de limite de polynomial.** Définit le degré maximal de régression polynomiale. Si vous définissez le degré maximal sur 1, l'algorithme de régression polynomiale est identique à l'algorithme de régression linéaire. Si vous spécifiez une valeur élevée pour le degré maximal de régression polynomiale, l'algorithme de régression polynomiale est enclin au surajustement. En d'autres termes, le modèle obtenu procède à une approximation précise des données

d'apprentissage ; toutefois, il échoue lorsqu'il est appliqué aux données non utilisées pour l'apprentissage.

**Utiliser la constante.** Si cette option est activée, elle oblige la courbe de régression à passer par l'origine. Cela signifie que le modèle ne contiendra pas de constante.

**Utiliser la sélection de fonction automatique.** Si cette option est activée et que vous ne précisez aucun niveau de signification minimal, l'algorithme tente de déterminer un sous-ensemble optimal des valeurs prédites possibles.

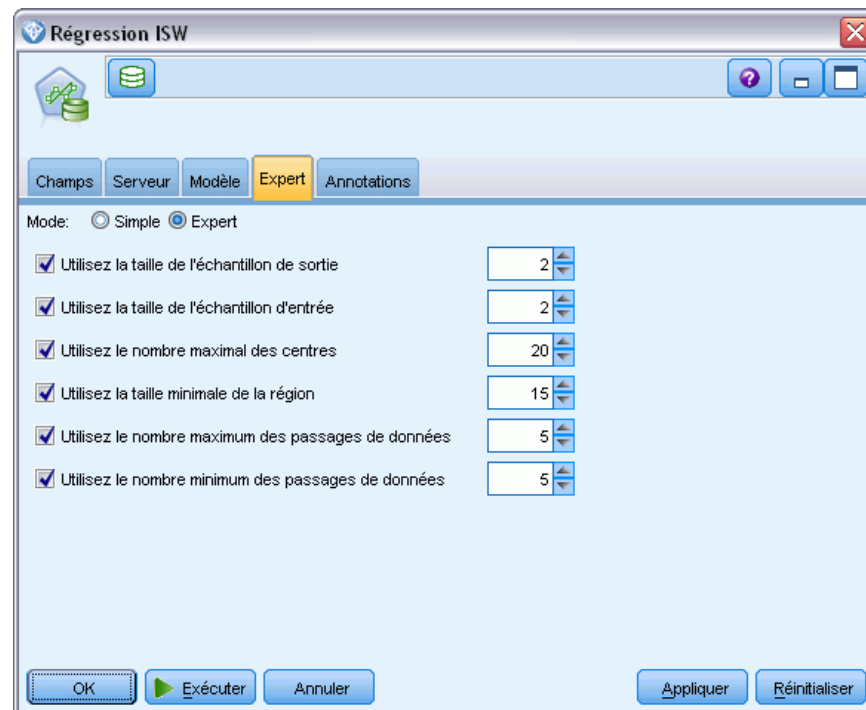
**Utiliser le niveau de signification minimal.** Lorsqu'un niveau de signification minimal est spécifié, la régression pas à pas est utilisée pour déterminer un sous-ensemble de valeurs prédites possibles. Seuls les champs indépendants dont la valeur de signification est supérieure à celle spécifiée contribuent au calcul du modèle de régression.

**Paramètres des champs.** Pour spécifier les options des champs d'entrée individuels, cliquez sur la ligne correspondante dans la colonne Paramètres du tableau Paramètres des champs et choisissez <Spécifier les paramètres>. [Pour plus d'informations, reportez-vous à la section Spécification des paramètres des champs de la régression sur p. 142.](#)

### Options Expert de la régression RBF

Figure 5-22

Noeud Régression ISW, Onglet Expert de la régression RBF



**Utiliser la taille de l'exemple de sortie.** Définit un exemple 1-dans-N pour la vérification et le test du modèle.

**Utiliser la taille de l'exemple d'entrée.** Définit un exemple 1-dans-N pour l'apprentissage.

**Utiliser le nombre de centre maximum.** Le nombre de centres maximum créés à chaque passage. Parce que le nombre de centres peut doubler par rapport au nombre d'origine lors d'un passage, le nombre de centres réel peut être supérieur au nombre spécifié.

**Utiliser la taille de région minimum.** Le nombre d'enregistrements minimum attribués à une région.

**Utiliser le nombre de passages de données maximum.** Le nombre de passages maximum dans les données d'entrée créées par l'algorithme. Si la valeur est spécifiée, elle doit être supérieure ou égale au nombre de passages minimum.

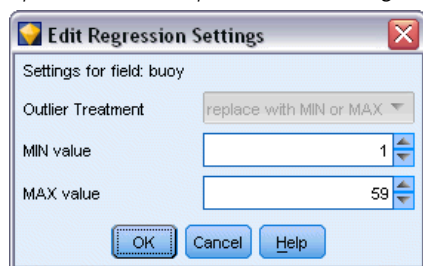
**Utiliser le nombre de passages de données minimum.** Le nombre de passages minimum dans les données d'entrée créées par l'algorithme. Spécifiez une valeur important uniquement si vous avez assez de données d'apprentissage et si vous êtes certain qu'un modèle adapté existe.

### **Spécification des paramètres des champs de la régression**

C'est ici que vous pouvez spécifier la plage de valeurs d'un champ d'entrée individuel.

Figure 5-23

Spécification des paramètres de régression d'un champ d'entrée



**Valeur MIN.** La valeur valide minimum de ce champ d'entrée.

**Valeur MAX.** La valeur valide maximum de ce champ d'entrée.

## **Classification ISW**

La fonction Mining de classification non supervisée recherche, dans les données d'entrée, les caractéristiques communes les plus fréquentes. Elle regroupe les données d'entrée en classes. Les membres de chaque classe sont dotés de propriétés similaires. Il n'y a pas de notion préconçue à propos des modèles existant dans les données. La classification non supervisée est un processus de découverte.

Le noeud Classification ISW vous propose les méthodes de classification suivantes :

- Démographique
- Kohonen
- BIRCH amélioré (Réduction itérative équilibrée et classification à l'aide de hiérarchies)



La technique de l'algorithme de **classification démographique** repose sur la proportion. Cette technique offre une classification non supervisée rapide et naturelle de très grandes bases de données. Le nombre de classes est choisi automatiquement (vous pouvez spécifier le nombre maximal de classes). Il existe un grand nombre de paramètres pouvant être configurés par l'utilisateur.

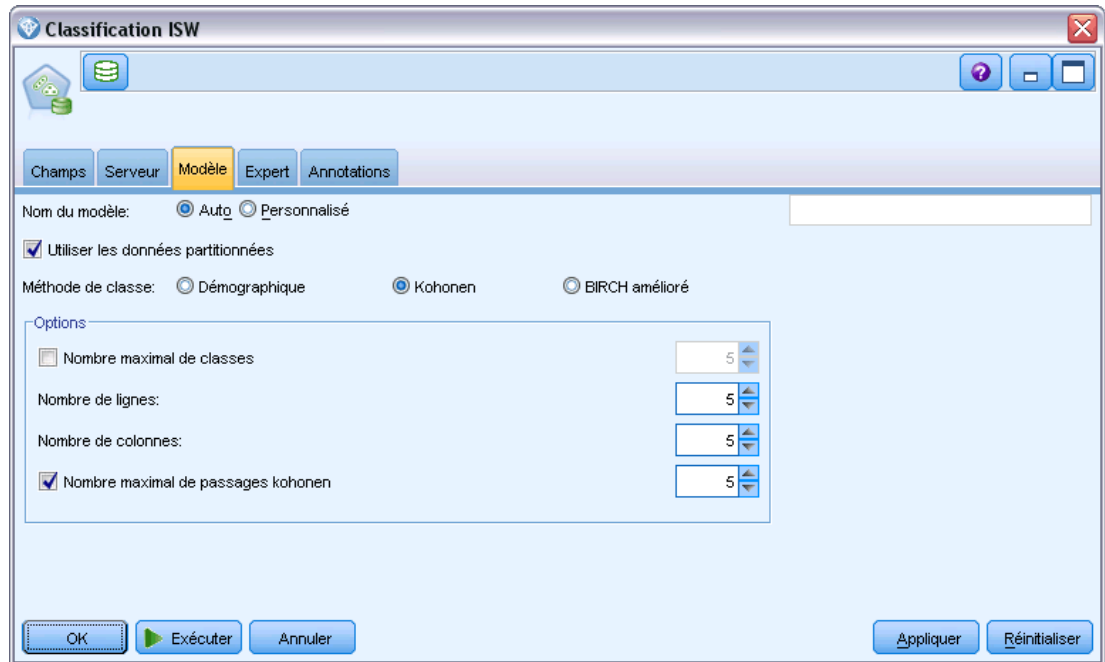
La technique de l'algorithme de **classification de Kohonen** repose sur le centre. La cartographie de Kohonen tente de placer les centres de classe aux positions qui minimisent la distance globale entre les enregistrements et ces mêmes centres. La séparabilité des classes n'est pas prise en compte. Les vecteurs de centre sont disposés sur une carte avec un certain nombre de colonnes et de lignes. Ils sont interconnectés de telle manière que le vecteur gagnant le plus proche d'un enregistrement d'apprentissage est ajusté, de même que les vecteurs dans son voisinage. Toutefois, plus les autres centres sont éloignés, moins ils sont ajustés.

La technique d'algorithme améliorée de **classification BIRCH** est basée sur la distribution et essaie de réduire la distance générale entre les enregistrements et leurs classes. La distance log-vraisemblance est utilisée par défaut pour déterminer la distance entre un enregistrement et une classe ; vous pouvez également sélectionner la distance euclidienne si tous les champs actifs sont des champs numériques. L'algorithme BIRCH effectue deux étapes indépendantes ; d'abord, il organise les enregistrements d'entrée dans un arbre Fonction de classification afin que les enregistrements similaires fassent partie des mêmes noeuds d'arbre puis il classe les feuilles de cet arbre dans la mémoire pour générer le résultat de classification final.

### ***Classification ISW - Options de modèle***

Dans l'onglet Modèle du noeud Classification, vous pouvez spécifier la méthode à utiliser pour créer des classes et également d'autres options associées.

Figure 5-24  
Noeud Classification ISW - Onglet Modèle



**Nom du modèle.** Vous pouvez générer le nom du modèle automatiquement sur la base du champ cible ou ID (ou du type de modèle si aucun de ces champs n'est spécifié) ou spécifier un nom personnalisé.

**Utiliser les données partitionnées.** Si un champ de partition est défini, cette option assure que seules les données provenant de la partition d'apprentissage sont utilisées pour générer le modèle. [Pour plus d'informations, reportez-vous à la section Noeud Partitionner dans le chapitre 4 dans Noeuds source, exécution et de sortie de IBM SPSS Modeler 15.](#)

**Méthode de classe.** Choisissez la méthode à utiliser pour créer les classes : Démographique, Kohonen ou BIRCH amélioré. [Pour plus d'informations, reportez-vous à la section Classification ISW sur p. 142.](#)

**Nombre maximal de classes.** La limitation du nombre de classes permet de gagner en temps d'exécution puisqu'elle empêche la création de nombreuses petites classes.

**Nombre de lignes / Nombre de colonnes.** (Méthode de Kohonen uniquement) Spécifie le nombre de lignes et de colonnes pour la cartographie de Kohonen. (Disponible uniquement si Nombre maximal de passages Kohonen est sélectionné et Nombre maximal de classes est désélectionné.)

**Nombre maximal de passages Kohonen.** (Méthode de Kohonen uniquement) Spécifie le nombre de passages de l'algorithme de classification sur les données durant les cycles d'apprentissage. Pour chaque passage, les vecteurs de centre sont ajustés afin de minimiser la distance totale entre les centres de classe et les enregistrements. De plus, le degré d'ajustement des vecteurs fait l'objet d'une diminution. Au premier passage, les ajustements sont grossiers. Au dernier passage, le degré d'ajustement des centres est plutôt réduit. Seuls les ajustements mineurs sont exécutés.

**Mesure de distance.** (Méthode BIRCH améliorée uniquement) Sélectionnez la mesure de la distance entre l'enregistrement et la classe utilisée par l'algorithme BIRCH. Vous pouvez choisir entre la distance log-vraisemblance qui est la distance par défaut ou la distance euclidienne.

*Remarque* : vous pouvez uniquement choisir la distance euclidienne lorsque tous les champs actifs sont des champs numériques.

**Nombre maximal de noeuds de feuilles.** (Méthode BIRCH améliorée uniquement) Le nombre maximal de noeuds de feuille que vous souhaitez sur l'arbre Fonction de classification. L'arbre Fonction de classification est le résultat de la première étape de l'algorithme BIRCH amélioré, dans lequel les enregistrements de données sont organisés dans un arbre de telle sorte que les enregistrements similaires sont rattachés au même noeud de feuille. La durée d'exécution de l'algorithme augmente avec le nombre de noeuds de feuille. La valeur par défaut est 1000.

**Passages BIRCH.** (Méthode BIRCH amélioré uniquement) Le nombre de passages de l'algorithme sur les données pour affiner les résultats de la classification. Le nombre de passages affecte la durée de traitement des exécutions d'apprentissage (chaque passage nécessitant une analyse complète des données) et la qualité du modèle. Les valeurs faibles entraînent une durée de traitement courte, toutefois, elles peuvent aussi donner des modèles de moindre qualité. Les valeurs élevées entraînent des durées de traitement plus longues et offrent généralement de meilleurs modèles. En moyenne, 3 passages ou plus donnent de bons résultats. La valeur par défaut est 3.

### ***Options expert de la classification ISW***

Dans l'onglet Expert du noeud Classification, vous pouvez indiquer des options avancées telles que des seuils de similarité, des limites de temps d'exécution et des pondérations de champ.

Figure 5-25  
Noeud Classification ISW - Onglet Expert

Mode:  Simple  Expert

Limite de la durée d'exécution

Limite de la durée d'exécution (minutes):

% minimal de données d'apprentissage:

Indiquer le seuil de similarité

Paramètres du champ :

Champ	Stockage	Paramètres
field2	Réal	
field3	Réal	
field4	Réal	
field5	Réal	
field6	Réal	
field7	Réal	

**Limite de la durée d'exécution.** Cochez cette case pour activer les options qui vous permettent de contrôler la durée nécessaire à la création du modèle. Vous pouvez spécifier une durée en minutes, un pourcentage minimal de données d'apprentissage à traiter ou les deux, et pour la méthode BIRCH, vous pouvez en plus spécifier le nombre maximal de noeuds de feuille à créer dans l'arbre FC.

**Indiquer le seuil de similarité.** (Classification démographique uniquement) La limite inférieure de la similarité de deux enregistrements de données qui appartiennent à la même classe. Par exemple, une valeur de 0,25 signifie que les enregistrements avec des valeurs ayant une similarité de 25% sont susceptibles d'être affectés à la même classe. Une valeur de 1,0 signifie que les enregistrements doivent être identiques pour apparaître dans la même classe.

**Paramètres des champs.** Pour spécifier les options des champs d'entrée individuels, cliquez sur la ligne correspondante dans la colonne Paramètres du tableau Paramètres des champs et choisissez <Spécifier les paramètres>.

### ***Spécification des paramètres des champs de la classification***

C'est ici que vous pouvez indiquer les options des champs d'entrée individuels.

Figure 5-26  
Spécification des paramètres de classe d'un champ d'entrée

**Modifier les paramètres de classification**

Paramètres du champ : field2

Pondération du champ: 1

Pondération de la valeur: aucun

Utiliser l'échelle de similarité: 1

Traitement des valeurs éloignées: aucun

Valeur MIN: 10 050

Valeur MAX: 60 081

OK Annuler Aide

**Pondération du champ.** Affecte une pondération plus ou moins importante au champ pendant le processus de création de modèle. Par exemple, si vous pensez que ce champ est relativement moins important pour le modèle que les autres champs, diminuez sa pondération par rapport aux autres champs.

**Pondération de la valeur.** Affecte une pondération plus ou moins importante aux valeurs particulières de ce champ. Certaines valeurs de champ peuvent être plus habituelles que d'autres valeurs. La coïncidence des valeurs rares dans un champ peut être plus importante pour une classe que la coïncidence des valeurs fréquentes. Vous pouvez choisir une des méthodes suivantes pour pondérer les valeurs de ce champ (dans les deux cas, les valeurs rares ont une pondération importante, alors que les valeurs habituelles ont une faible pondération).

- **Logarithmique.** Affecte une pondération à chaque valeur en fonction du logarithme de sa probabilité dans les données d'entrée.
- **Probabiliste.** Affecte une pondération à chaque valeur en fonction de sa probabilité dans les données d'entrée.

Pour chaque méthode, vous pouvez également choisir une option avec compensation pour compenser la pondération de valeur appliquée à chaque champ. Si vous compensez la pondération de valeur, l'importance générale du champ pondéré est égale à celle d'un champ non pondéré. Cela fonctionne ainsi, quel que soit le nombre de valeurs possibles. La pondération compensée affecte uniquement l'importance relative des coïncidences dans l'ensemble des valeurs pondérées.

**Utiliser l'échelle de similarité.** Cochez cette case si vous souhaitez utiliser une échelle de similarité pour contrôler le calcul des mesures de similarité d'un champ. Spécifiez l'échelle de similarité sous la forme d'un nombre absolu. Cette spécification est prise en compte uniquement pour les champs numériques actifs. Si vous n'indiquez pas d'échelle de similarité, la valeur par défaut (moitié de l'écart-type) est utilisée. Pour obtenir un grand nombre de classes, diminuez la similarité moyenne entre les paires de classes avec des échelles de similarité plus petites pour les champs numériques.

**Traitement des valeurs éloignées.** Les valeurs éloignées sont des valeurs de champ qui se trouvent en-dehors de la plage de valeurs spécifiées pour ce champ, comme défini par la valeur MIN et la valeur MAX. Vous pouvez choisir la façon de traiter les valeurs éloignées de ce champ.

- Le paramètre par défaut, aucun, signifie qu'aucune action spécifique n'est entreprise pour les valeurs éloignées.
- Si vous choisissez remplacer par MIN ou MAX, une valeur de champ inférieure à la valeur MIN ou supérieure à la valeur MAX est remplacée par les valeurs de MIN ou MAX selon le cas. Dans ce cas, vous pouvez définir les valeurs de MIN et MAX.
- Si vous choisissez traiter comme manquantes, les valeurs éloignées sont traitées comme des valeurs manquantes et ignorées. Dans ce cas, vous pouvez définir les valeurs de MIN et MAX.

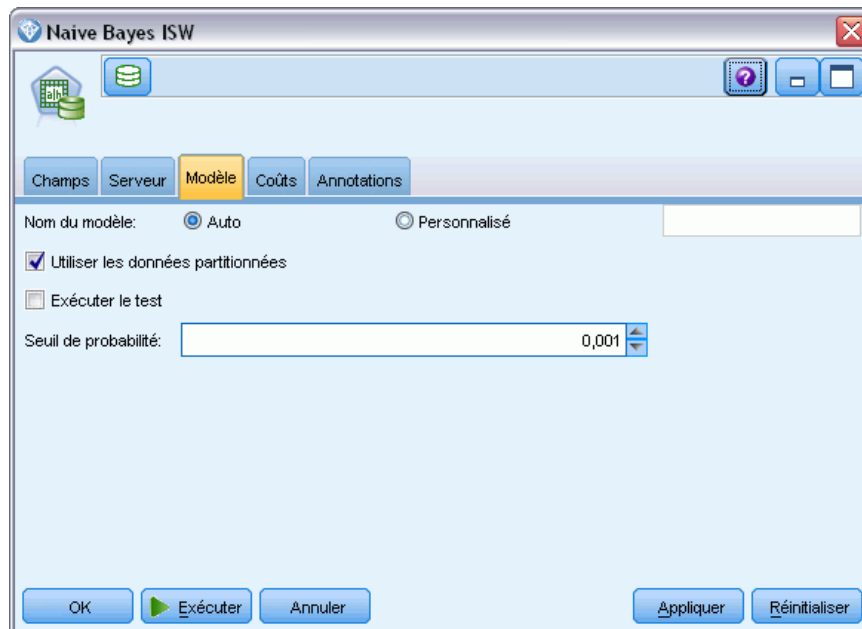
## Naive Bayes ISW

Naive Bayes est un algorithme bien connu pour les problèmes de classification. Ce modèle est appelé *naïve*, car il considère toutes les variables de prévision proposées comme étant indépendantes les unes des autres. Naive Bayes est un algorithme rapide et évolutif qui calcule les probabilités conditionnelles des combinaisons d'attributs et de l'attribut cible. Une probabilité indépendante est établie à partir des données d'apprentissage. Cette probabilité détermine la vraisemblance de chaque classe cible et est calculée en fonction de l'occurrence de chaque catégorie de valeur issue des variables d'entrée.

L'algorithme de classification Naive Bayes ISW est un classificateur probabiliste. Il se base sur les modèles de probabilité qui intègrent des hypothèses d'indépendance importantes.

### Options des modèles Naive Bayes d'ISW

Figure 5-27  
Onglet Modèle du noeud Naive Bayes ISW



**Nom du modèle.** Vous pouvez générer le nom du modèle automatiquement sur la base du champ cible ou ID (ou du type de modèle si aucun de ces champs n'est spécifié) ou spécifier un nom personnalisé.

**Utiliser les données partitionnées.** Si un champ de partition est défini, cette option assure que seules les données provenant de la partition d'apprentissage sont utilisées pour générer le modèle. [Pour plus d'informations, reportez-vous à la section Noeud Partitionner dans le chapitre 4 dans Noeuds source, exécution et de sortie de IBM SPSS Modeler 15.](#)

**Exécuter le test.** Vous pouvez réaliser un test. Un test IBM InfoSphere Warehouse Data Mining est alors exécuté une fois le modèle créé sur la partition d'apprentissage. Cela engendre un passage sur la partition test, permettant d'établir des informations sur la qualité du modèle, des graphiques Lift, etc.

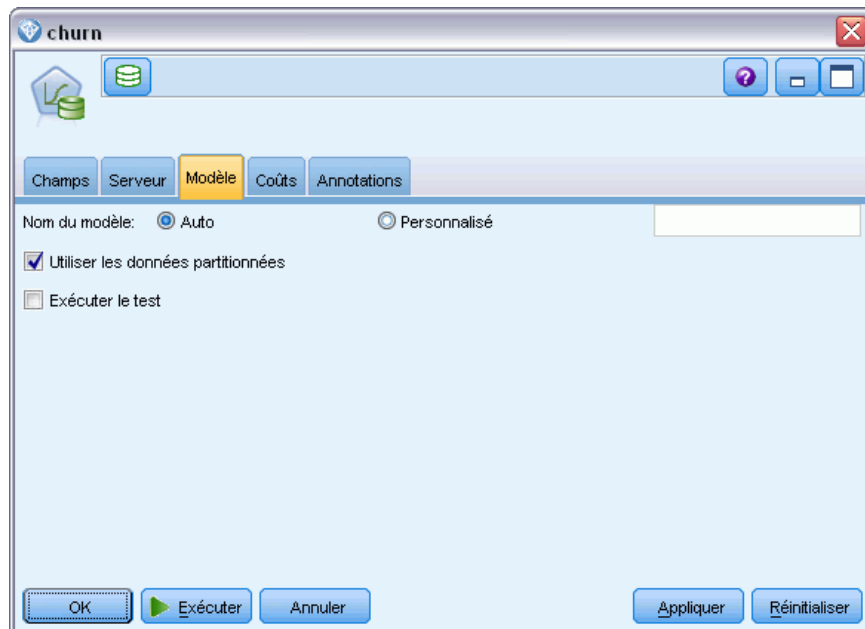
**Seuil de probabilité.** Le seuil de probabilité définit une probabilité pour toutes les combinaisons des variables indépendantes et des valeurs cible qui n'apparaissent pas dans les données d'apprentissage. Cette probabilité doit être comprise entre 0 et 1. La valeur par défaut est 0,001.

## Régression logistique ISW

La régression logistique, également appelée régression nominale, est une technique statistique permettant de classer des enregistrements en fonction des valeurs de leurs champs d'entrée. Il s'agit d'une régression analogue à linéaire, mais l'algorithme Régression logistique ISW utilise un champ cible booléen (binaire) au lieu d'un champ numérique.

### Régression logistique d'ISW - Options de modèle

Figure 5-28  
Noeud Régression logistique ISW - Onglet Modèle



**Nom du modèle.** Vous pouvez générer le nom du modèle automatiquement sur la base du champ cible ou ID (ou du type de modèle si aucun de ces champs n'est spécifié) ou spécifier un nom personnalisé.

**Utiliser les données partitionnées.** Si un champ de partition est défini, cette option assure que seules les données provenant de la partition d'apprentissage sont utilisées pour générer le modèle. [Pour plus d'informations, reportez-vous à la section Noeud Partitionner dans le chapitre 4 dans Noeuds source, exécution et de sortie de IBM SPSS Modeler 15.](#)

**Exécuter le test.** Vous pouvez réaliser un test. Un test IBM InfoSphere Warehouse Data Mining est alors exécuté une fois le modèle créé sur la partition d'apprentissage. Cela engendre un passage sur la partition test, permettant d'établir des informations sur la qualité du modèle, des graphiques Lift, etc.

## **Séries temporelles ISW**

Les algorithmes Séries temporelles ISW vous permettent de prédire des événements futurs en fonction d'événements connus du passé.

Comme les méthodes de régression standard, les algorithmes Séries temporelles prédisent une valeur numérique. Contrairement aux méthodes de régression standard, les prévisions de séries temporelles se concentrent sur les valeurs futures d'une série ordonnée. .

Les algorithmes de séries temporelles sont des algorithmes univariés. Cela signifie que la variable indépendante est une colonne de temps ou d'ordre. Les prévisions sont basées sur des valeurs passées. Elles ne sont pas basées sur d'autres colonnes indépendantes.

Les algorithmes de séries temporelles sont différents des algorithmes de régression commune en ce qu'ils ne prédisent pas uniquement des valeurs futures mais intègrent également des cycles saisonniers dans leurs prévisions.

La fonction de Data Mining des séries temporelles propose les algorithmes suivants permettant de prévoir les tendances à venir :

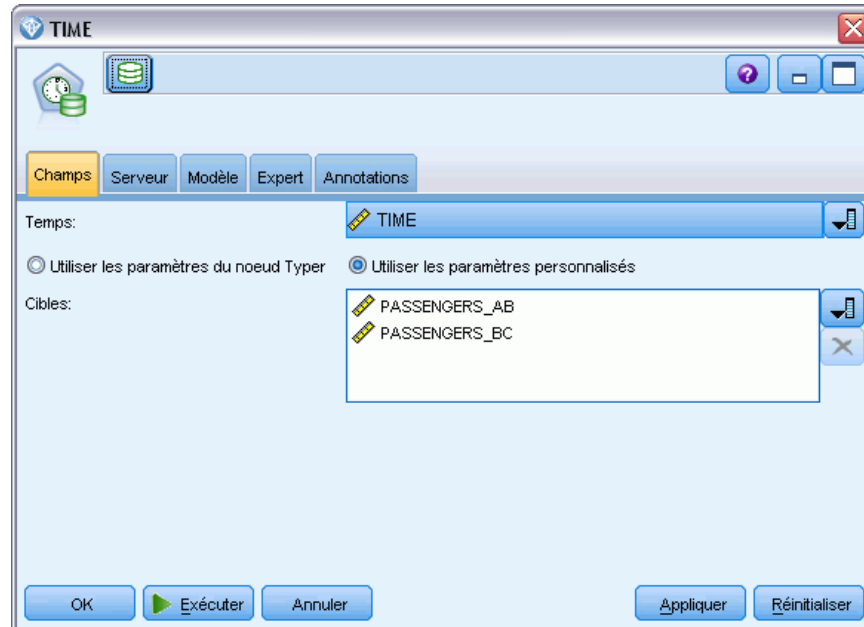
- AutoRegressive Integrated Moving Average (processus autorégressif moyenne mobile intégré - ARIMA).
- Lissage exponentiel
- Décomposition de tendance saisonnière

L'algorithme qui crée les meilleures prévisions pour vos données dépend des différentes hypothèses de modèles. Vous pouvez calculer toutes les prévisions en même temps. Les algorithmes calculent des prévisions détaillées qui incluent le comportement saisonnier des séries temporelles d'origine. Si IBM InfoSphere Warehouse est installé, vous pouvez utiliser le visualiseur de séries temporelles pour évaluer et comparer les courbes de résultats.



## Options des champs des séries temporelles ISW

Figure 5-29  
Onglet Champs du nœud Séries temporelles ISW



**Heure :** Sélectionnez le champ d'entrée qui contient la série temporelle. Il doit s'agir d'un champ avec un type de stockage Date, Heure, Horodatage, Réel ou Entier.

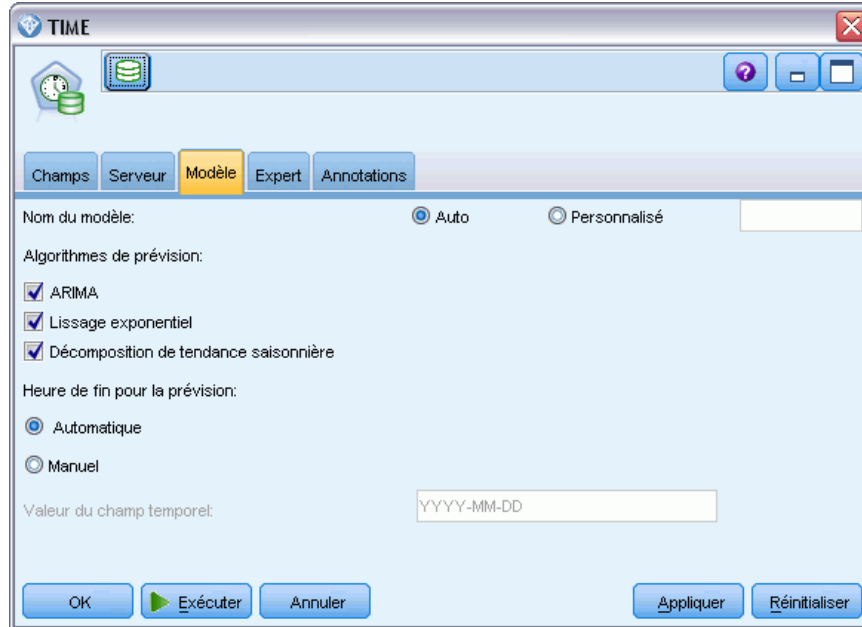
**Utiliser les paramètres du nœud Typier.** Cette option indique au nœud d'utiliser les informations du champ à partir d'un nœud Typier en amont. Il s'agit de la valeur par défaut.

**Utiliser les paramètres personnalisés.** Cette option indique au nœud d'utiliser les informations du champ spécifiées ici au lieu des informations données dans un nœud Typier en amont. Une fois cette option sélectionnée, renseignez les champs ci-dessous.

**Cibles.** Sélectionnez un ou plusieurs champs cible. Cela revient à définir le rôle du champ sur la valeur *Cible* dans un nœud Typier.

## Options du modèle de séries temporelles ISW

Figure 5-30  
Onglet Modèle du noeud Séries temporelles ISW



**Nom du modèle.** Vous pouvez générer le nom du modèle automatiquement sur la base du champ cible ou ID (ou du type de modèle si aucun de ces champs n'est spécifié) ou spécifier un nom personnalisé.

**Algorithmes de prévision.** Sélectionnez les algorithmes à utiliser pour la modélisation. Vous pouvez choisir un ou plusieurs des éléments suivants :

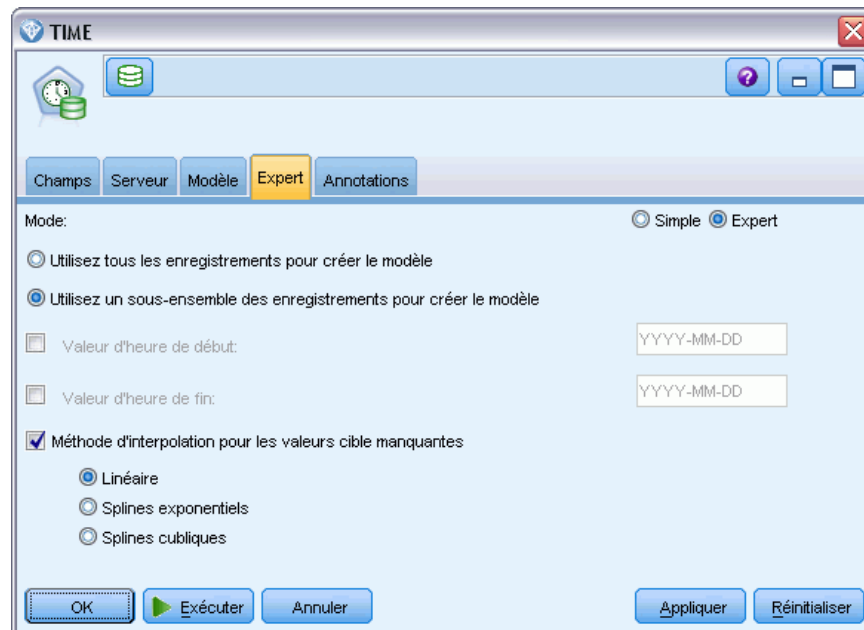
- ARIMA
- Lissage exponentiel
- Décomposition des tendances saisonnières.

**Heure de fin des prévisions.** Spécifiez si l'heure de fin des prévisions doit être calculée automatiquement ou spécifiée manuellement.

**Valeur du champ Temps.** Lorsque l'heure de fin des prévisions est définie manuellement, saisissez l'heure de fin des prévisions. La valeur à saisir dépend du type du champ Temps ; par exemple, si le type est un entier représentant les heures, alors vous pouvez entrer 48 pour arrêter les prévisions après que des données aient été traitées pendant 48 heures. Ce champ peut également vous demander d'entrer une date ou une heure comme valeur finale.

## Options de l'expert de séries temporelles ISW

Figure 5-31  
Onglet Expert du noeud Séries temporelles ISW



**Utiliser tous les enregistrements pour créer le modèle.** Il s'agit du paramètre par défaut ; tous les enregistrements sont analysés lorsque le modèle est créé.

**Utiliser un sous-ensemble des enregistrements pour créer le modèle.** Sélectionnez cette option pour créer le modèle à partir d'une partie des données disponibles. Par exemple, cette option peut être utile lorsque vous disposez d'une quantité excessive de données semblables.

Saisissez la valeur de l'heure de début et la valeur de l'heure de fin pour identifier les données à utiliser. Veuillez noter que les valeurs que vous pouvez entrer dans ces champs dépendent du type du champ heure ; par exemple, il peut s'agir d'un certain nombre d'heures ou de jours, ou de dates ou d'heures spécifiques.

**Méthode d'interpolation pour les valeurs cibles manquantes.** Si vous traitez des données contenant une ou plusieurs valeurs manquantes, sélectionnez la méthode à utiliser pour les calculer. Vous pouvez choisir l'un des éléments suivants :

- Linéaire
- Splines exponentiels
- Splines cubiques

## Affichage des modèles de séries temporelles ISW

Les modèles des séries temporelles ISW sont des résultats sous la forme de modèle brut, qui contient des informations extraites des données mais qui n'est pas conçu pour générer directement des prévisions.

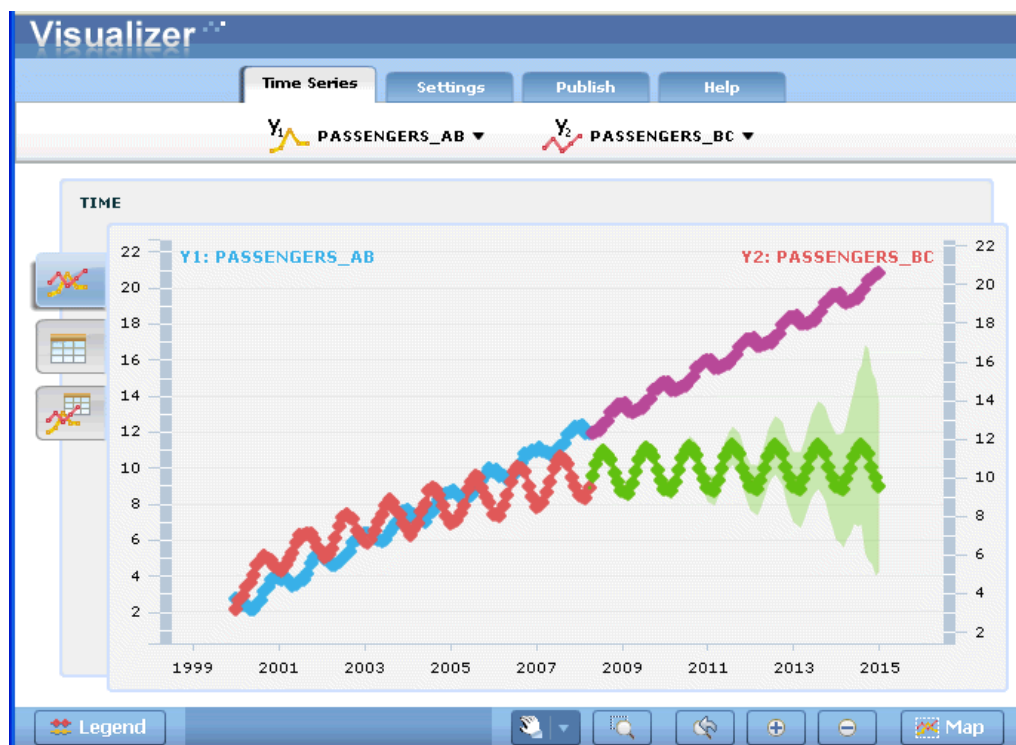
Figure 5-32  
Icône de modèle brut



Pour plus d'informations, reportez-vous à la section Modèles bruts dans le chapitre 3 dans *Noeuds de modélisation de IBM SPSS Modeler 15*.

Si le client IBM InfoSphere Warehouse est installé, vous pouvez utiliser l'outil Visualiseur de séries temporelles pour un affichage graphique de vos données de séries temporelles.

Figure 5-33  
Modèle des séries temporelles ISW affichées dans le visualiseur



Utilisation de l'outil Visualiseur de séries temporelles :

- ▶ Vérifiez que vous avez terminé les tâches d'intégration de IBM® SPSS® Modeler à IBM InfoSphere Warehouse. [Pour plus d'informations, reportez-vous à la section Activation de l'intégration à IBM InfoSphere Warehouse sur p. 108.](#)
- ▶ Faites un double clic sur l'icône du modèle brut dans la palette Modèles.

- Dans l'onglet Serveur de la boîte de dialogue, cliquez sur le bouton Affichage pour afficher le visualiseur dans votre navigateur par défaut.

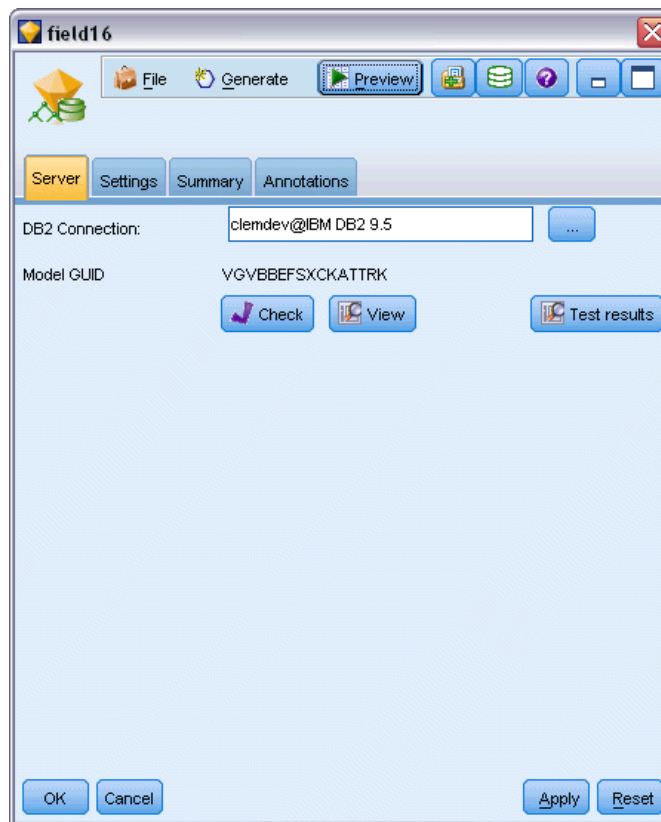
## Nuggets de modèle ISW Data Mining

Vous pouvez créer des modèles à partir des noeuds Arbre décision, Association, Séquence, Régression et Classification non supervisée ISW inclus dans IBM® SPSS® Modeler.

### Nugget de modèle ISW - Onglet Serveur

L'onglet Serveur fournit des options permettant d'exécuter des vérifications de cohérence et de lancer l'outil IBM Visualizer.

Figure 5-34  
Nugget de modèle ISW - Onglet Serveur



IBM® SPSS® Modeler peut vérifier la cohérence en stockant une chaîne clé modèle générée identique dans les modèles SPSS Modeler et ISW. Pour exécuter la vérification de la cohérence, cliquez sur le bouton Vérifier de l'onglet Serveur. [Pour plus d'informations, reportez-vous à la section Gestion des modèles DB2 sur p. 117.](#)

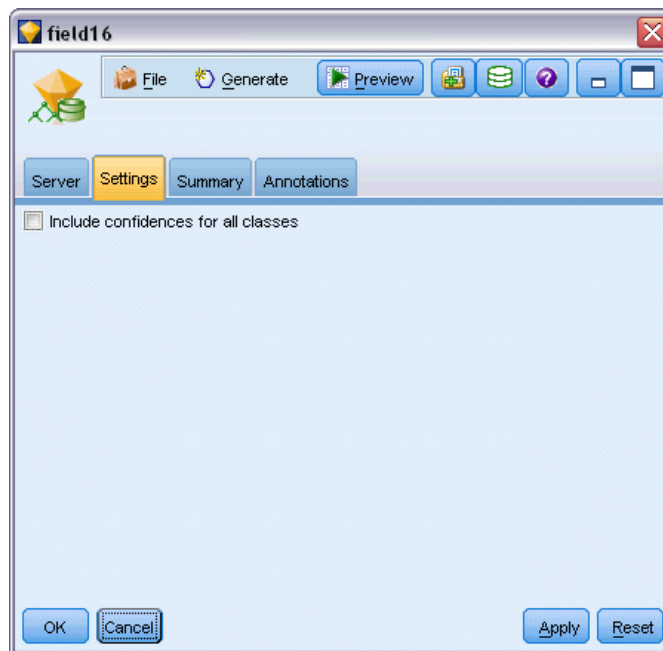
L'outil Visualizer est la seule méthode permettant de parcourir les modèles InfoSphere Warehouse Data Mining. L'outil peut être installé en option avec InfoSphere Warehouse Data Mining. [Pour plus d'informations, reportez-vous à la section Activation de l'intégration à IBM InfoSphere Warehouse sur p. 108.](#)

- Cliquez sur **Affichage** pour lancer l'outil de visualisation. L'affiche de l'outil dépend du type de noeud généré. L'outil de visualisation peut, par exemple, afficher les classes prédites s'il est lancé à partir d'un nugget de modèle d'arbre de décision ISW.
- Cliquez sur **Résultats de test** (arbres décision et séquence uniquement) pour lancer l'outil de visualisation et afficher la qualité générale du modèle généré.

### ***Nugget de modèle ISW - Onglet Paramètres***

Dans IBM® SPSS® Modeler, en général, une seule prévision et une probabilité ou confiance associées sont livrées. De plus, une option utilisateur permettant d'afficher les probabilités pour chaque résultat (identique à celle trouvée dans la régression logistique) correspond à une option de temps de score disponible sur l'onglet Paramètres du nugget de modèle.

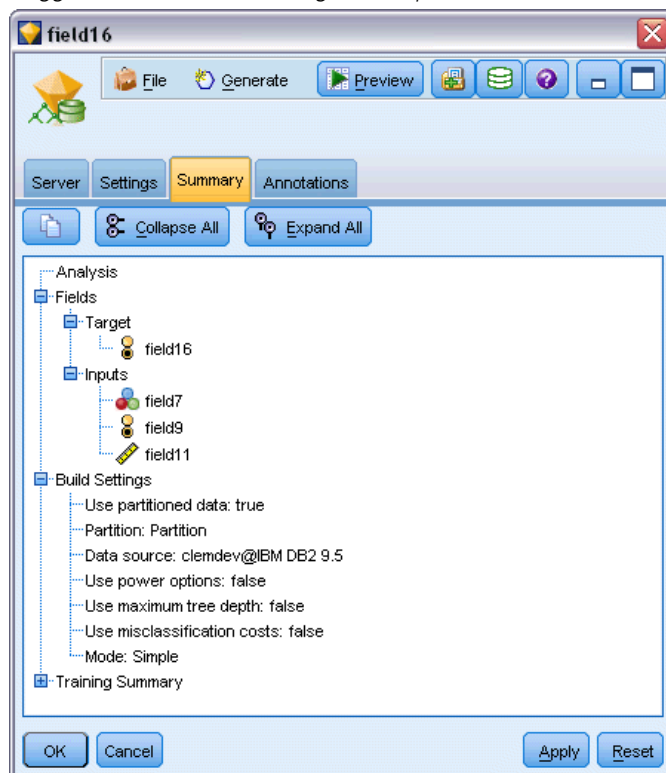
Figure 5-35  
*Nugget de modèle ISW - Onglet Paramètres*



**Inclure les confiances de toutes les classes.** Pour chaque revenu possible du champ cible, ajoutez une colonne donnant le niveau de confiance.

## Nugget de modèle ISW - Onglet Récapitulatif

Figure 5-36  
Nugget de modèle ISW - Onglet Récapitulatif



L'onglet Récapitulatif d'un nugget de modèle contient des informations sur le modèle lui-même (*Analyse*), sur les champs utilisés dans le modèle (*Champs*), sur les paramètres utilisés pour la construction du modèle (*Créer des paramètres*), ainsi que sur l'apprentissage du modèle (*Récapitulatif de l'apprentissage*).

Lorsque vous accédez au noeud pour la première fois, l'arborescence des résultats de l'onglet Récapitulatif est réduite. Pour afficher les résultats qui vous intéressent, utilisez la commande à gauche d'un élément pour le développer ou cliquez sur le bouton Développer tout pour afficher tous les résultats. Pour masquer les résultats lorsque vous avez terminé de les consulter, utilisez la commande de développement pour réduire les résultats voulus ou cliquez sur le bouton Réduire tout pour réduire tous les résultats.

**Analyse.** Affiche des informations sur le modèle en question. Si vous avez exécuté un noeud Analyse relié à ce nugget de modèle, les informations issues de l'analyse figureront également dans cette section. [Pour plus d'informations, reportez-vous à la section Noeud Analyse dans le chapitre 6 dans \*Noeuds source, exécution et de sortie de IBM SPSS Modeler 15\*.](#)

**Champs.** Répertorie les champs utilisés comme cibles et entrées lors de la création du modèle.

**Créer des paramètres.** Contient des informations sur les paramètres utilisés lors de la création du modèle.

**Récapitulatif de l'apprentissage.** Indique le type du modèle, le flux utilisé pour le créer, l'utilisateur qui l'a créé, ainsi que sa date et sa durée de création.

## Exemples d'ISW Data mining

IBM® SPSS® Modeler pour Windows est livré avec des démonstrations de flux illustrant le processus d'exploration de base de données. Ces flux sont disponibles dans le dossier d'installation de IBM® SPSS® Modeler à l'emplacement suivant :

`\Demos\Database_Modeling\IBM DB2 ISW`

*Remarque :* Le dossier Démonstrations est accessible à partir du groupe de programmes SPSS Modeler du menu Démarrer de Windows.

Les flux suivants peuvent être utilisés consécutivement les uns après les autres comme exemple du processus d'exploration de base de données :

- *1\_upload\_data.str*—Utilisé pour nettoyer et envoyer des données à partir d'un fichier plat vers DB2.
- *2\_explore\_data.str*—Utilisé comme exemple d'exploration de données avec SPSS Modeler.
- *3\_build\_model.str*—Utilisé pour créer un modèle d'arbre décision ISW.
- *4\_evaluate\_model.str*—Utilisé comme exemple d'évaluation de modèle avec SPSS Modeler.
- *5\_deploy\_model.str*—Utilisé pour déployer le modèle de détermination des scores dans la base de données.

L'ensemble de données utilisé dans les exemples de flux concerne les applications pour carte de crédit et présente un problème de classification avec un mélange de variables indépendantes catégorielles et continues. Pour plus d'informations sur cet ensemble de données, reportez-vous au fichier suivant installé avec SPSS Modeler :

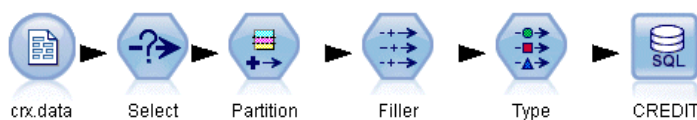
`\Demos\Database_Modeling\IBM DB2 ISW\crx.names`

Cet ensemble de données est disponible à partir du référentiel d'apprentissage automatique UCI <http://archive.ics.uci.edu/ml/>.

### Exemple de flux : Envoi de données

Le premier exemple de flux, *1\_upload\_data.str*, est utilisé pour nettoyer et envoyer les données à partir d'un fichier plat vers DB2.

Figure 5-37  
Exemple de flux utilisé pour envoyer les données



Le nœud Remplacer est utilisé pour le traitement des valeurs manquantes et remplace les champs vides lus dans le fichier texte *crx.data* par des valeurs *NULL*.



## Exemple de flux : Exploration des données

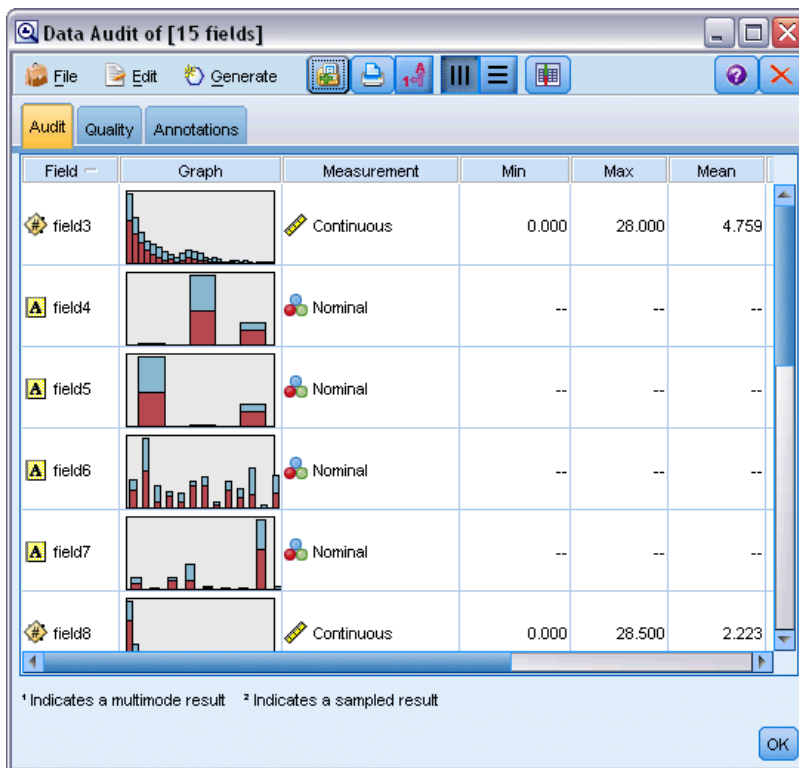
Le deuxième exemple de flux, *2\_explore\_data.str*, est utilisé pour décrire l'exploration de données dans IBM® SPSS® Modeler.

Figure 5-38  
Exemple de flux utilisé pour explorer les données



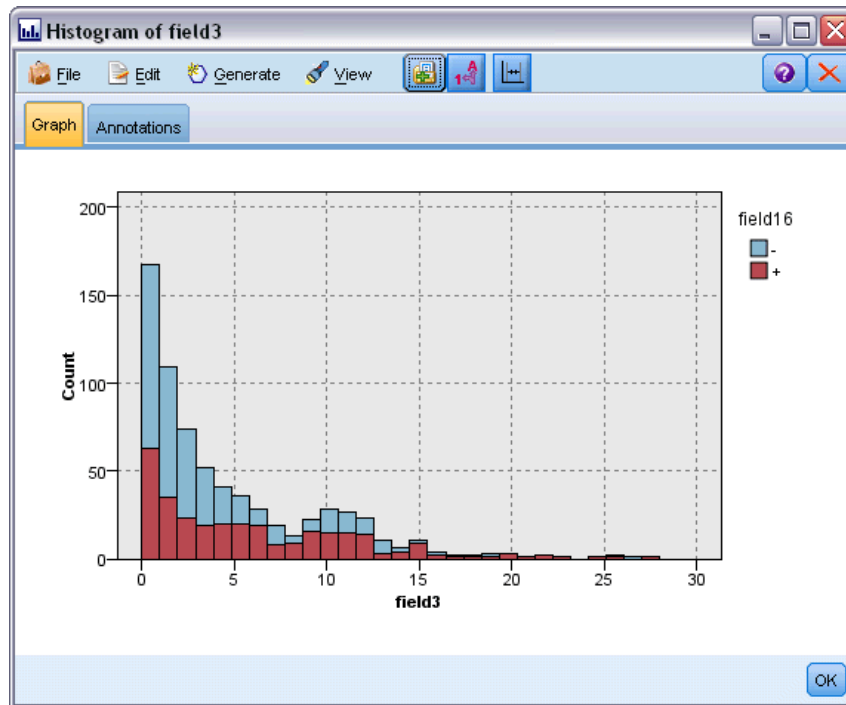
Lors de l'exploration de données, l'une des étapes standard consiste à relier un noeud Audit données aux données. Le noeud Audit données est disponible dans la palette de noeuds Sortie.

Figure 5-39  
Résultats de l'audit de données



Vous pouvez utiliser la sortie d'un noeud Audit données pour obtenir un aperçu général des champs et de la proportion des données. Pour obtenir un graphique plus détaillé permettant d'explorer un champ de manière plus précise, double-cliquez sur le graphique voulu dans la fenêtre Audit données.

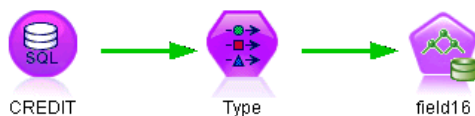
Figure 5-40  
Histogramme créé en double-cliquant sur un graphique dans la fenêtre Audit données



### Exemple de flux : Créer un modèle

Le troisième exemple de flux, *3\_build\_model.str*, illustre la création de modèles dans IBM® SPSS® Modeler. Vous pouvez relier le noeud de modélisation de base de données au flux et double-cliquer dessus pour définir les paramètres de création.

Figure 5-41  
Exemple de flux de modélisation de base de données, où les noeuds en violet font référence à l'exécution d'opérations dans une base de données

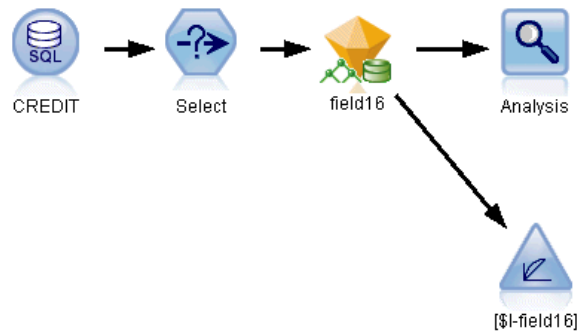


Les onglets *Modèle* et *Expert* du noeud de modélisation permettent d'ajuster la profondeur d'arbre maximale et d'arrêter la division d'un noeud à partir de la création de l'arbre décision d'origine en définissant la pureté maximale et le nombre minimal d'observations par noeud interne. [Pour plus d'informations, reportez-vous à la section Arbre décision ISW sur p. 123.](#)

### Exemple de flux : Evaluation d'un modèle

Le quatrième exemple de flux, *4\_evaluate\_model.str*, illustre les avantages que présente l'utilisation de IBM® SPSS® Modeler pour la modélisation dans la base de données. Une fois le modèle exécuté, vous pouvez l'ajouter à nouveau à votre flux de données et l'évaluer à l'aide des divers outils proposés par SPSS Modeler.

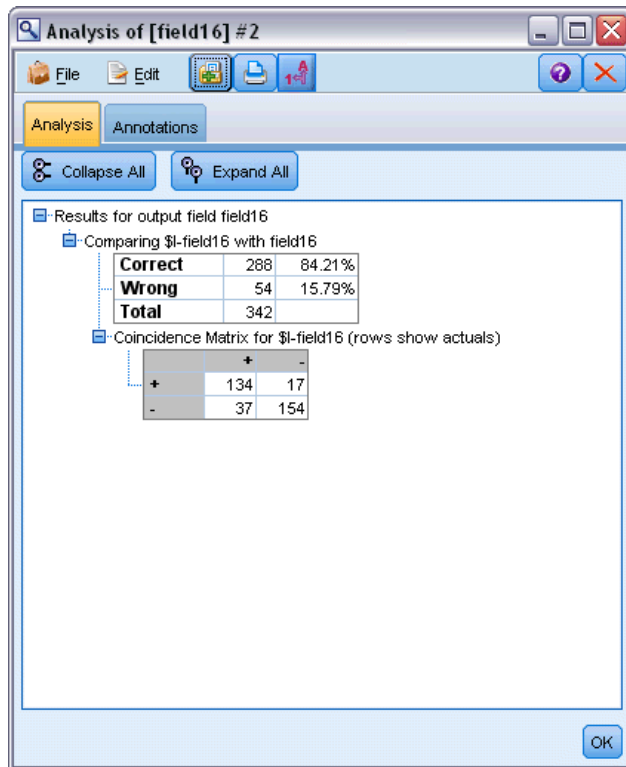
Figure 5-42  
Exemple de flux utilisé pour l'évaluation du modèle



La première fois que vous ouvrez le flux, le nugget de modèle (*field16*) n'est pas inclus dans le flux. Ouvrez le noeud source CREDIT et vérifiez que vous avez spécifié une source de données. Puis, à condition que vous ayez exécuté le flux *3\_build\_model.str* pour créer un nugget *field16* dans la palette Modèles, vous pouvez exécuter les noeuds déconnectés en cliquant sur le bouton Exécuter de la barre d'outils (le bouton avec un triangle vert). Cela exécute un script qui copie le nugget *field16* dans le flux, le connecte aux noeuds existants puis exécute les noeuds terminaux dans le flux.

Vous pouvez relier un noeud Analyse (disponible sur la palette Sortie) pour créer une matrice de coïncidences illustrant le motif des correspondances entre chaque champ (prédit) généré et son champ cible. Exécutez le noeud Analyse pour afficher les résultats.

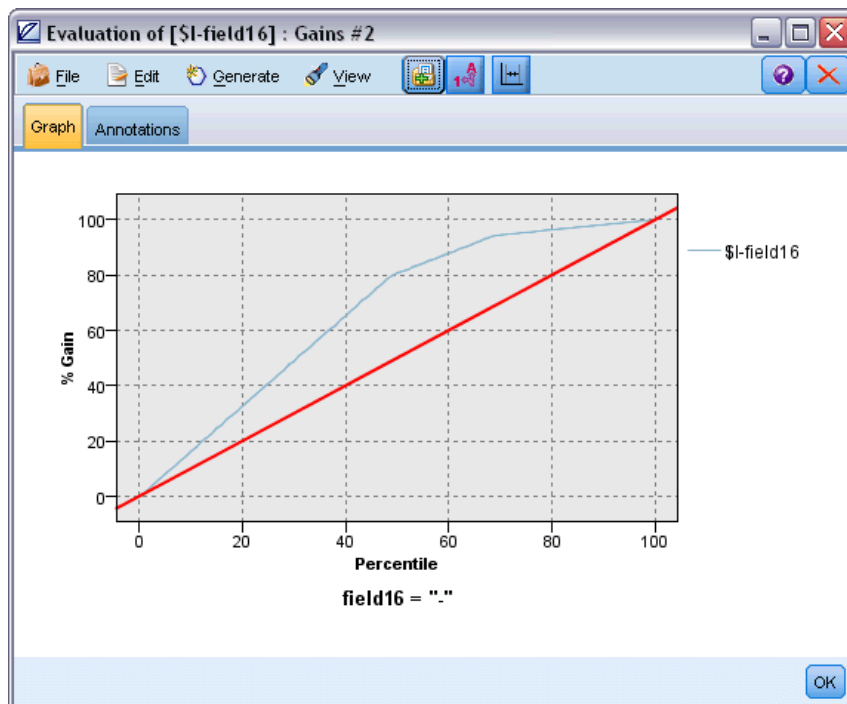
Figure 5-43  
Résultats du noeud Analyse



Le tableau créé indique que 84,21% des prévisions générées par l’algorithme Arbre décision ISW sont correctes.

Vous pouvez également créer un graphique de gains pour afficher les améliorations de précision apportées par le modèle. Reliez un noeud Evaluation au modèle généré, puis exécutez le flux pour afficher les résultats.

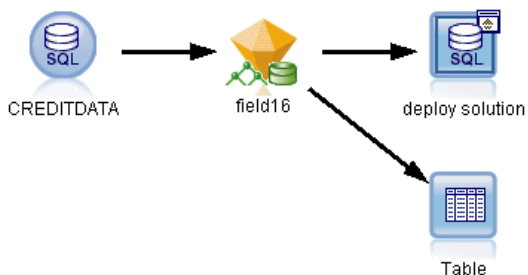
Figure 5-44  
Graphique Gains généré à l'aide du noeud Evaluation



### Exemple de flux : Déploiement d'un modèle

Une fois que la précision du modèle vous convient, vous pouvez le déployer afin de l'utiliser avec des applications externes ou pour écrire des scores dans la base de données. Dans l'exemple de flux *5\_deploy\_model.str*, les données sont lues à partir de la table CREDIT. Lorsque le noeud d'exportation de la base de données *solution de déploiement* est exécuté, les données ne sont pas réellement évaluées. En effet, le flux crée le fichier image publié *credit\_scorer.pim* et le fichier de paramètres publié *credit\_scorer.par*.

Figure 5-45  
Exemple de flux utilisé pour déployer le modèle



Comme dans l'exemple précédent, le flux exécute un script qui copie le nugget *field16* dans le flux de la palette Modèles, le connecte aux noeuds existants puis exécute les noeuds terminaux dans le flux. Dans ce cas, vous devez d'abord spécifier une source de données à la fois dans la source de la base de données et dans les noeuds d'exportation.

# ***Modélisation de la base de données avec IBM Netezza Analytics***

## ***IBM SPSS Modeler et IBM Netezza Analytics***

IBM® SPSS® Modeler prend en charge l'intégration à IBM® Netezza® Analytics qui permet d'effectuer des analyses avancées sur les serveurs IBM Netezza. Ces fonctions sont accessibles via l'interface utilisateur graphique et l'environnement de développement orienté workflow de SPSS Modeler, ce qui permet aux clients d'utiliser les algorithmes de Data mining directement dans l'environnement Netezza d'IBM.

SPSS Modeler prend en charge l'intégration des algorithmes de Netezza Analytics suivants.

- Arbres décision
- K-means
- Bayes Net
- Naive Bayes
- KNN
- Classification par division
- ACP
- Arbre de régression
- Régression linéaire

Pour plus d'informations sur les algorithmes, consultez le *Guide du développeur Netezza Analytics* et le *Guide de référence Netezza Analytics*.

## ***Conditions requises pour l'intégration à IBM Netezza Analytics***

Les conditions préalables suivantes sont obligatoires pour réaliser une modélisation dans la base de données via IBM® Netezza® Analytics. N'hésitez pas à contacter l'administrateur de base de données pour vous assurer que ces conditions sont remplies.

- IBM® SPSS® Modeler s'exécutant en mode local ou en fonction d'une installation IBM® SPSS® Modeler Server sur Windows ou Unix (à l'exception de zLinux, pour lequel les pilotes ODBC Netezza d'IBM ne sont pas disponibles).
- IBM Netezza Performance Server 6.0 ou version ultérieure exécutant le package IBM® SPSS® In-Database Analytics.

- Une source de données ODBC pour se connecter à une base de données IBM Netezza. [Pour plus d'informations, reportez-vous à la section Activation de l'intégration à IBM Netezza Analytics sur p. 166.](#)
- Génération et optimisation SQL activées dans SPSS Modeler. [Pour plus d'informations, reportez-vous à la section Activation de l'intégration à IBM Netezza Analytics sur p. 166.](#)

*Remarque* : La modélisation de base de données et l'optimisation SQL requièrent l'activation de la connectivité à SPSS Modeler Server sur l'ordinateur IBM® SPSS® Modeler. Avec ce paramètre activé, vous pouvez accéder aux algorithmes de la base de données, effectuer le push back de SQL directement depuis SPSS Modeler et accéder à SPSS Modeler Server. Pour vérifier l'état actuel de la licence, choisissez ce qui suit dans le menu SPSS Modeler.

Aide > A propos de > Informations supplémentaires

Si la connectivité est activée, vous voyez l'option Activation du serveur dans l'onglet Etat de la licence.

[Pour plus d'informations, reportez-vous à la section Connexion au IBM SPSS Modeler Server dans le chapitre 3 dans \*Guide de l'utilisateur de IBM SPSS Modeler 15.\*](#)

## **Activation de l'intégration à IBM Netezza Analytics**

L'activation de l'intégration avec IBM® Netezza® Analytics comprend les étapes suivantes :

- Configuration de Netezza Analytics
- Création d'une source de données ODBC
- Activation de l'intégration dans IBM® SPSS® Modeler
- Activation de la génération SQL et de l'optimisation dans SPSS Modeler

Elles sont décrites dans les sections suivantes.

### **Configuration de IBM Netezza Analytics**

Pour installer et configurer IBM® Netezza® Analytics, consultez la documentation de Netezza Analytics —en particulier le *Guide d'installation Netezza Analytics* —pour des informations supplémentaires. La section *Définition des droits de la base de données* de ce guide contient des détails sur les scripts qui doivent être exécutés pour autoriser les flux IBM® SPSS® Modeler à écrire dans la base de données.

*Remarque* : Si vous allez utiliser des nœuds basés sur un calcul matriciel (ACP Netezza et régression linéaire Netezza), le moteur de matrice Netezza doit être initialisé en exécutant `CALL NZM..INITIALIZE()`, sinon l'exécution des procédures stockées échouera. L'initialisation est une étape d'installation unique pour chaque base de données.

### **Création d'une source de données ODBC pour IBM Netezza Analytics**

Pour activer la connexion entre la base de données IBM Netezza et IBM® SPSS® Modeler, vous devez créer un nom de source de données ODBC (DSN).



Avant de créer un DSN, vous devez avoir des connaissances de base des sources de données et des pilotes ODBC, ainsi que de la prise en charge de la base de données dans SPSS Modeler. [Pour plus d'informations, reportez-vous à la section Accès aux données dans le chapitre 2 dans Guide d'administration et des performances de IBM SPSS Modeler Server 15.](#)

Si vous exécutez le système en mode réparti sur le IBM® SPSS® Modeler Server, créez le DSN sur l'ordinateur serveur. Si vous exécutez le système en mode (client) local, créez le DSN sur l'ordinateur client.

### **Clients Windows**

- ▶ A partir du CD de votre *client Netezza*, exécutez le fichier *nzodbcsetup.exe* pour démarrer le programme d'installation. Suivez les instructions à l'écran pour installer le pilote. Pour obtenir des instructions complètes, consultez le *Guide d'installation et de configuration de IBM Netezza ODBC, JDBC et OLE DB*.

- ▶ Créez le DSN (nom de source de données).

*Remarque* : La séquence de menus dépend de la version de Windows que vous utilisez.

- **Windows XP.** Dans le menu Démarrer, sélectionnez Panneau de configuration. Double-cliquez sur Outils d'administration, puis sur Sources de données (ODBC).
  - **Windows Vista.** Dans le menu Démarrer, sélectionnez Panneau de configuration; puis Maintenance du système. Double-cliquez sur Outils d'administration, sélectionnez Sources de données (ODBC), puis cliquez sur Ouvrir.
  - **Windows 7.** Dans le menu Démarrer, sélectionnez Panneau de configuration; puis Sécurité du système, puis Outils d'administration. Sélectionnez Sources de données (ODBC), puis cliquez sur Ouvrir.
- ▶ Cliquez sur l'onglet du DSN système, puis sur Ajouter.
  - ▶ Sélectionnez NetezzaSQL dans la liste et cliquez sur Finish.
  - ▶ Sur l'onglet DSN Options de l'écran de configuration du pilote Netezza ODBC, entrez le nom de la source de données de votre choix, le nom d'hôte ou l'adresse IP du serveur IBM Netezza, le numéro de port de la connexion, la base de données de l'instance IBM Netezza que vous utilisez, ainsi que les détails de votre nom d'utilisateur et de votre mot de passe pour la connexion de base de données. Cliquez sur le bouton Help pour obtenir une explication sur les champs.
  - ▶ Cliquez sur le bouton Test Connection et vérifiez que vous pouvez vous connecter à la base de données.
  - ▶ Une fois que vous êtes connectés correctement, cliquez sur OK à plusieurs reprises pour quitter l'écran de l'administrateur de source de données ODBC.

### **Serveurs Windows**

La procédure pour Windows Server est la même que la procédure client pour Windows XP.

### **Serveurs UNIX ou Linux**

La procédure suivante s'applique aux serveurs UNIX ou Linux (à l'exception de zLinux, pour lequel les pilotes ODBC IBM Netezza ne sont pas disponibles).

- ▶ À partir du CD de votre *client Netezza*, copiez le fichier `<platform>cli.package.tar.gz` approprié vers un emplacement temporaire sur le serveur.
- ▶ Extrayez le contenu de l'archive à l'aide des commandes `gunzip` et `untar`.
- ▶ Ajoutez des permissions d'exécution au script *décompacter* qui est extrait.
- ▶ Exécutez le script en répondant aux invites à l'écran.
- ▶ Modifiez le fichier *modelersrv.sh* pour inclure les lignes suivantes.

```
./usr/IBM/SPSS/SDAP61_notfinal/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/SDAP61_notfinal; export NZ_ODBC_INI_PATH
```

- ▶ Recherchez le fichier `/usr/local/nz/lib64/odbc.ini` et copiez son contenu dans le fichier *odbc.ini* installé avec SDAP 6.1 (celui défini par la variable d'environnement \$ODBCINI).

*Remarque* : pour les systèmes Linux 64 bits, le paramètre *Pilote* désigne par erreur le pilote 32 bits. Lorsque vous copiez le contenu de *odbc.ini* à l'étape précédente, modifiez le chemin dans ce paramètre en fonction, par exemple :

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

- ▶ Modifiez les paramètres dans la définition du DSN Netezza afin d'indiquer la base de données à utiliser.
- ▶ Redémarrez le SPSS Modeler Server et testez l'utilisation des nœuds d'exploration de la base de données Netezza sur le client.

### **Activation de l'intégration de IBM Netezza Analytics dans IBM SPSS Modeler**

- ▶ Dans le menu principal de IBM® SPSS® Modeler, sélectionnez Outils > Options > Programmes externes.
- ▶ Cliquez sur l'onglet IBM Netezza.

**Activer l'intégration de Netezza Data Mining.** Active la palette Modélisation de la base de données (si elle n'est pas déjà visible) en bas de la fenêtre SPSS Modeler et ajoutez les nœuds des algorithmes Netezza Data Mining.

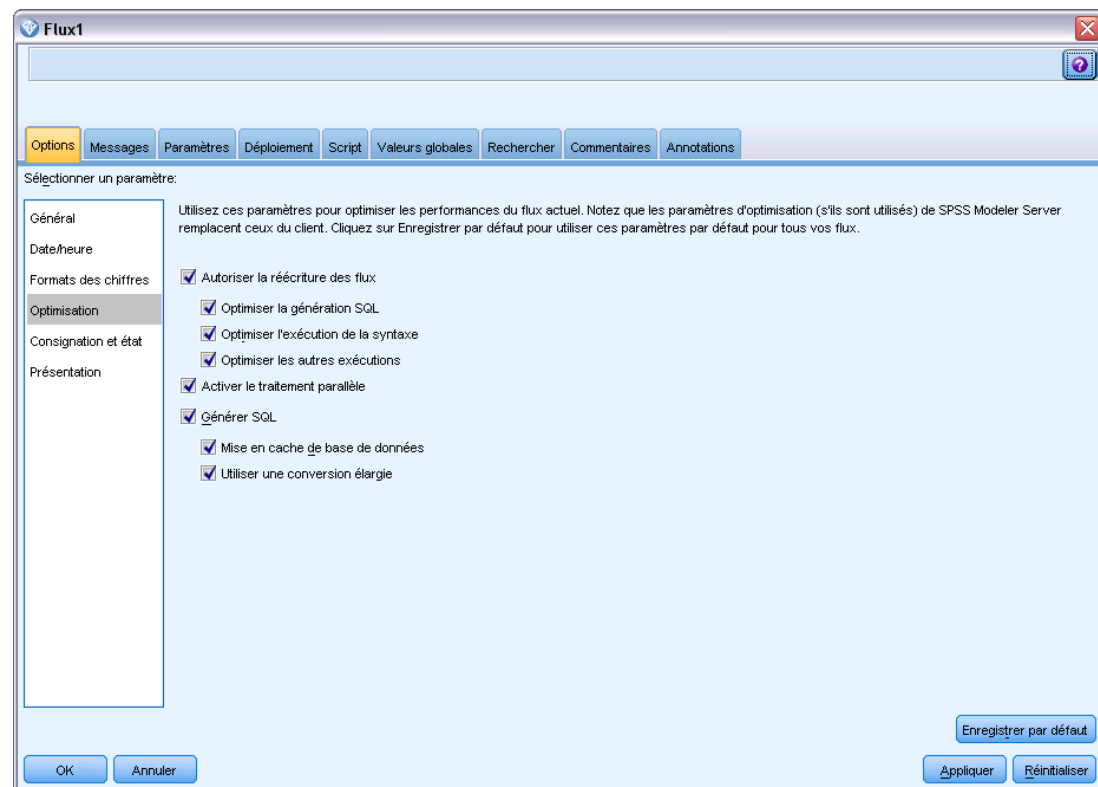
**Connexion Netezza.** Cliquez sur le bouton Modifier et choisissez la chaîne de connexion Netezza que vous avez configurée précédemment lors de la création de la source ODBC. [Pour plus d'informations, reportez-vous à la section Création d'une source de données ODBC pour IBM Netezza Analytics sur p. 166.](#)

## Activation de la génération SQL et de l'optimisation

Parce qu'il est probable que vous utilisiez de très grands ensembles de données, nous vous conseillons d'activer les options de génération et d'optimisation SQL dans IBM® SPSS® Modeler pour un fonctionnement optimisé.

- Dans les menus SPSS Modeler, sélectionnez :  
Outils > Propriétés du flux > Options

Figure 6-1  
Paramètres d'optimisation



- Cliquez sur l'option Optimisation dans le volet de navigation.
- Confirmez que l'option Générer SQL est bien activée. Ce paramètre doit être utilisé pour que la modélisation de base de données puisse fonctionner.
- Sélectionnez Optimiser la génération SQL et Optimiser les autres exécutions (cette opération n'est pas forcément nécessaire, mais elle est vivement recommandée pour optimiser les performances).

Pour plus d'informations, reportez-vous à la section Définition des options d'optimisation pour les flux dans le chapitre 5 dans *Guide de l'utilisateur de IBM SPSS Modeler 15*.

## Création de modèles avec IBM Netezza Analytics

Chacun des algorithmes pris en charge a un nœud de modélisation correspondant. Vous pouvez accéder aux nœuds de modélisation IBM Netezza à partir de l'onglet Modélisation de la base de données dans la palette des nœuds. [Pour plus d'informations, reportez-vous à la section Palette de noeuds dans le chapitre 3 dans \*Guide de l'utilisateur de IBM SPSS Modeler 15\*.](#)

### Remarques sur les données

Les champs, dans la source de données, peuvent contenir des variables de divers types de données en fonction du nœud de modélisation. Dans IBM® SPSS® Modeler, les types de données sont nommés **niveaux de mesure**. L'onglet Champs du nœud de modélisation utilise des icônes pour indiquer les types de niveaux de mesure autorisés pour ses champs d'entrée et cible. [Pour plus d'informations, reportez-vous à la section Niveaux de mesure dans le chapitre 4 dans \*Noeuds source, exécution et de sortie de IBM SPSS Modeler 15\*.](#)

**Champ cible.** Le champ cible est celui dont vous essayez de prévoir la valeur. Quand une cible peut être spécifiée, un seul des champs de données source peut être sélectionné comme champ cible.

**Champ ID d'enregistrement.** Indique le champ permettant d'identifier chaque observation de façon unique. Par exemple, il peut s'agir d'un champ d'ID, comme *ID client*. Si les données sources ne comportent pas de champ ID, vous pouvez créer ce champ à l'aide d'un nœud Calculer, comme le montre la procédure suivante.

- ▶ Sélectionnez le nœud source.
- ▶ Dans l'onglet Ops sur champs de la palette de nœuds, double-cliquez sur le noeud Calculer.
- ▶ Ouvrez le nœud Calculer en double-cliquant sur son icône sur l'espace de travail.
- ▶ Dans le champ Calculer champ, saisissez (par exemple) ID.
- ▶ Dans le champ Formule, saisissez @INDEX et cliquez sur OK.
- ▶ Connectez le nœud Calculer au reste du flux.

### Traitement des valeurs nulles

Si les données d'entrée contiennent des valeurs nulles, l'utilisation de nœuds Netezza peut provoquer des messages d'erreur ou des flux longue durée, nous vous recommandons donc de supprimer les enregistrements contenant des valeurs nulles. Utilisez la méthode suivante.

- ▶ Reliez un nœud Sélectionner au nœud source.
- ▶ Définissez l'option Mode du nœud Sélectionner sur Supprimer.
- ▶ Saisissez ce qui suit dans le champ Condition :  
`@NULL(champ1) [ou @NULL(champ2)[... ou @NULL(champM)]`
- Assurez-vous de saisir chaque champ d'entrée.
- ▶ Connectez le nœud Sélectionner au reste du flux.

### Sortie de modèle

Il est possible qu'un flux contenant un nœud de modélisation Netezza produise des résultats légèrement différents chaque fois qu'il est exécuté. Ceci est dû au fait que l'ordre dans lequel le nœud lit les données source n'est pas toujours le même, car les données sont lues dans des tables temporaires avant la création du modèle. Toutefois, de telles différences sont négligeables.

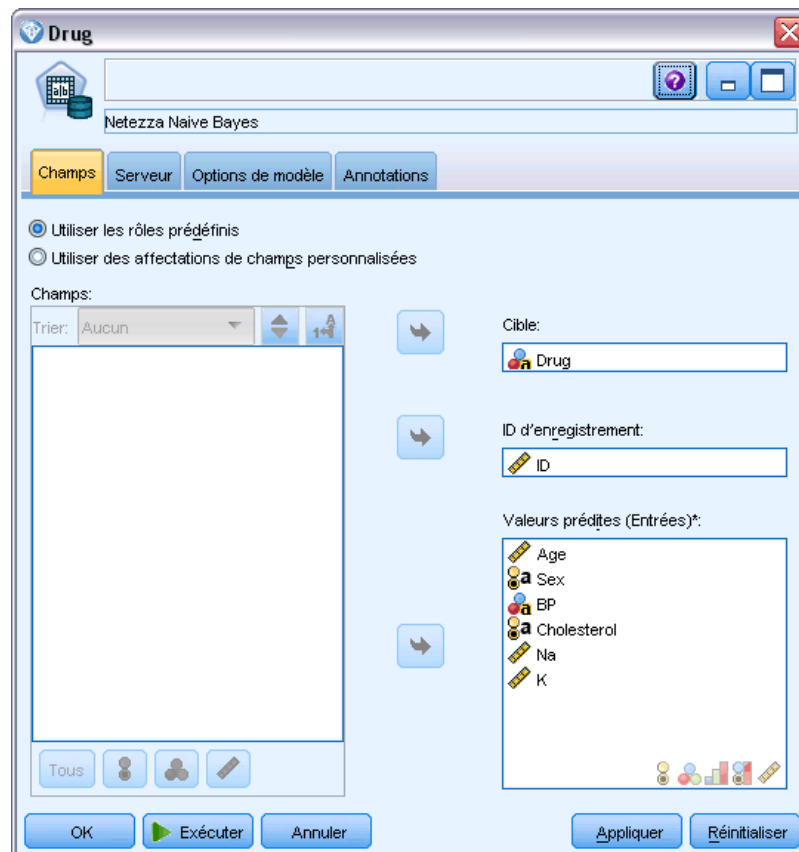
### Commentaires généraux

- Dans IBM® SPSS® Collaboration and Deployment Services, il est impossible de créer des configurations de scoring à l'aide des flux contenant les nœuds de modélisation de la base de données IBM Netezza.
- L'exportation ou l'importation PMML est impossible pour les modèles créés par des nœuds Netezza.

## Modèles Netezza - Options de champs

Dans l'onglet Champs, vous pouvez choisir d'utiliser les paramètres du rôle de champ déjà définis dans les noeuds en amont, ou de réaliser manuellement des affectations de champs.

Figure 6-2  
Exemples d'options de champs Netezza



**Utiliser des rôles prédéfinis.** Cette option utilise les paramètres de rôle (cibles, variables indépendantes, etc.) pour un nœud Type en amont (ou l'onglet Types d'un nœud source en amont). [Pour plus d'informations, reportez-vous à la section Définition du rôle du champ dans le chapitre 4 dans \*Noeuds source, exécution et de sortie de IBM SPSS Modeler 15\*.](#)

**Utiliser des affectations de champs personnalisées.** Sélectionnez cette option si vous souhaitez affecter manuellement des cibles, des variables indépendantes et d'autres rôles sur cet écran.

**Champs.** Utilisez les boutons fléchés pour affecter manuellement des éléments à partir de cette liste aux divers champs de rôle à droite de l'écran. Les icônes indiquent les niveaux de mesure valides pour chaque champ de rôle.

Cliquez sur le bouton Tous pour sélectionner tous les champs de la liste ou cliquez sur un bouton de niveau de mesure individuelle pour sélectionner tous les champs avec un niveau de mesure.

**Cible.** Sélectionnez un champ comme cible pour la prédiction. Pour les modèles linéaires généralisés, consultez également le champ Essais sur cet écran.

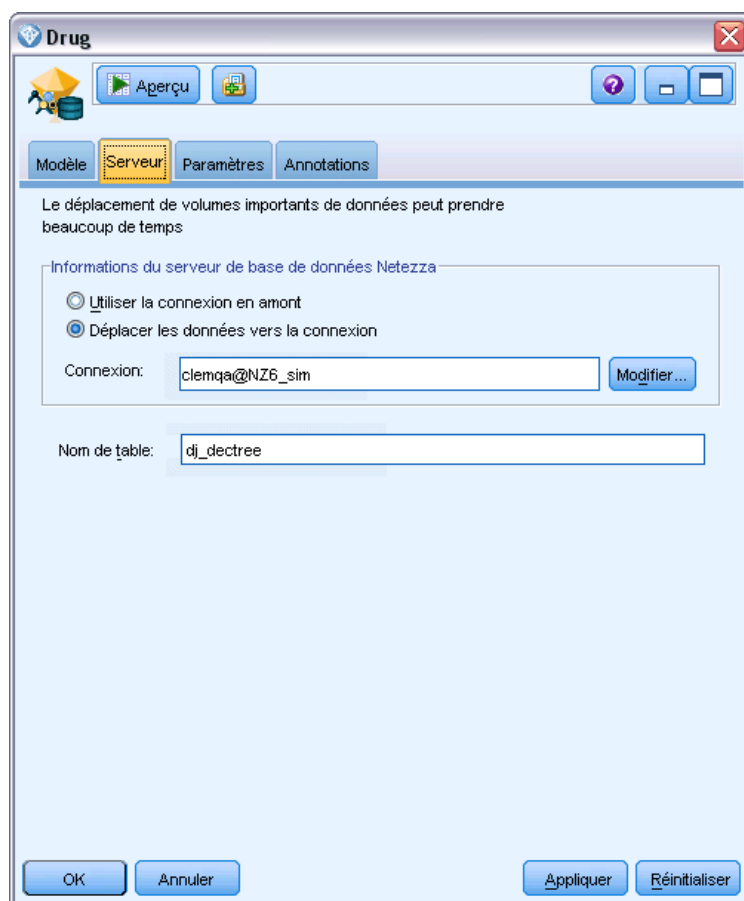
**ID d'enregistrement.** Le champ est utilisé comme identificateur d'enregistrement unique.

**Valeurs prédites (Entrées).** Choisissez un ou plusieurs champs comme entrées pour la prédiction.

### ***Modèles Netezza - Options du serveur***

Dans cet onglet, spécifiez la base de données IBM Netezza où le modèle doit être stocké.

Figure 6-3  
Exemples d'options du serveur Netezza



**Informations sur le serveur Netezza DB.** C'est ici que vous spécifiez les informations de connexion pour la base de données à utiliser pour le modèle.

- **Utiliser la connexion en amont.** (par défaut) Utilise les informations de connexion spécifiées dans un noeud en amont, par exemple le noeud source Base de données. *Remarque* : cette option fonctionne uniquement si tous les noeuds en amont peuvent utiliser les répercussions SQL. Dans ce cas, il est inutile de déplacer les données en-dehors de la base de données car le SQL implémente entièrement tous les noeuds en amont.
- **Déplacer les données vers la connexion.** Déplace les données vers la base de données spécifiée ici. Ceci permet de faire fonctionner la modélisation si les données se trouvent dans une autre base de données IBM Netazza ou une base de données d'un autre fabricant, ou même si les données se trouvent dans un fichier plat. De plus, les données sont de nouveau déplacées vers la base de données spécifiée ici si les données ont été extraites parce qu'un noeud n'a pas effectué de répercussions SQL. Cliquez sur le bouton Edit pour rechercher et sélectionner une connexion. *Avertissement* : IBM® Netezza® Analytics est généralement utilisé avec de très grands ensembles de données. Le transfert de grandes quantités de données entre des bases de données, ou hors d'une base de données puis dans cette même base de données, peut prendre beaucoup de temps et doit être évité aussi souvent que possible.

**Nom de la table.** Le nom de la table de base de données où le modèle doit être stocké. *Remarque :* Cela doit être une nouvelle table; vous ne pouvez pas utiliser une table existante pour cette opération.

#### **Commentaires**

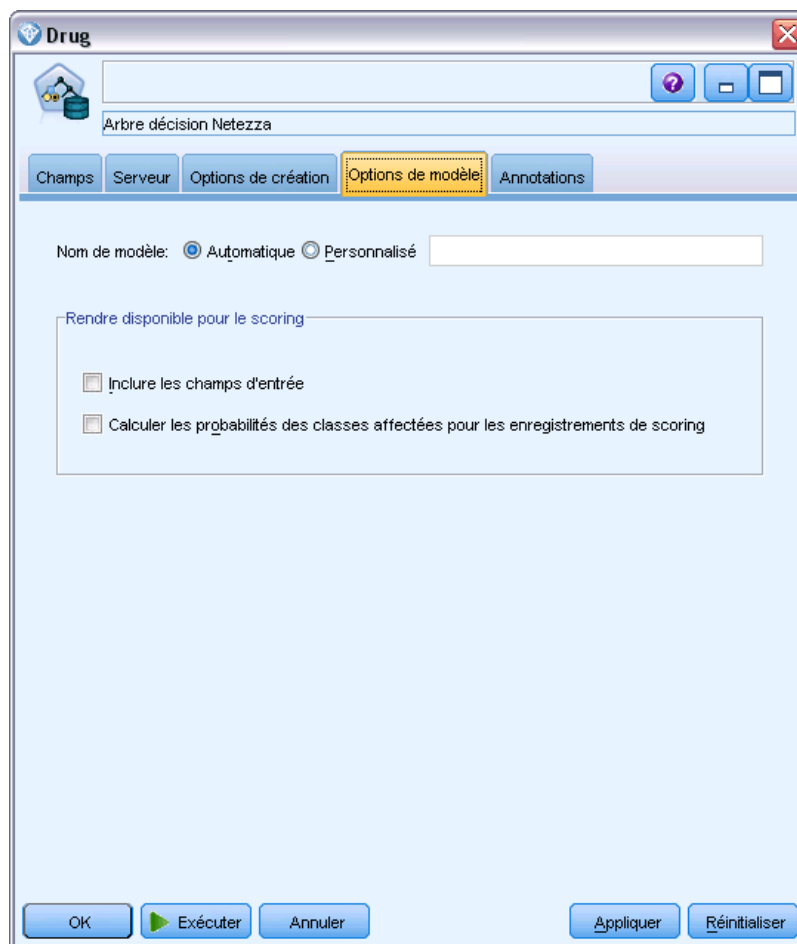
- La connexion utilisée pour la modélisation n'est pas tenue d'être identique à la connexion utilisée dans le nœud source d'un flux. Par exemple, vous pouvez utiliser un flux qui accède aux données d'une base de données IBM Netezza, télécharge ces données vers IBM® SPSS® Modeler pour qu'elles soient nettoyées ou fassent l'objet de diverses manipulations, puis télécharge ces données à une autre base de données IBM Netezza pour la modélisation. Cependant, veuillez noter qu'une telle configuration peut avoir un impact négatif sur les performances.
- Le nom de la source de données ODBC est incorporé de manière efficace dans chaque flux SPSS Modeler. Si un flux créé sur un hôte est exécuté sur un autre hôte, le nom de la source de données doit être identique sur chaque hôte. Vous pouvez également sélectionner une autre source de données dans l'onglet Serveur de chaque source ou nœud de modélisation.

### **Modèles Netezza - Options du modèle**

Dans l'onglet Options de modèle, vous pouvez choisir de spécifier un nom pour le modèle ou de générer automatiquement un nom. Vous pouvez également définir les valeurs par défaut pour les options de scoring.



Figure 6-4  
Exemples d'options de modèle Netezza



**Nom du modèle.** Vous pouvez générer le nom du modèle automatiquement sur la base du champ cible ou ID (ou du type de modèle si aucun de ces champs n'est spécifié) ou spécifier un nom personnalisé.

**Rendre disponible pour le scoring.** Vous pouvez définir ici les valeurs par défaut pour les options de scoring qui apparaissent dans la boîte de dialogue du nugget de modèle. Pour les détails des options, consultez la rubrique d'aide de l'onglet Paramètres de ce nugget particulier.

## Arbres décision Netezza

Un arbre de décision est une structure hiérarchique qui représente un modèle de classification. Avec un modèle d'arbre décision, vous pouvez développer un système de classification qui prédit ou classe les observations futures à partir d'un ensemble de données d'apprentissage. La classification prend la forme d'une arborescence dans laquelle les branches représentent des points de division de la classification. La division décompose de manière récursive les données en sous-groupes, jusqu'à ce qu'un point d'arrêt soit atteint. Les nœuds d'arbre aux points d'arrêt sont

appelés des **feuilles**. Chaque feuille affecte une étiquette, connue comme **étiquette de classe**, pour les membres de son sous-groupe, ou de sa classe.

La sortie du modèle prend la forme d'une représentation textuelle de l'arbre. Chaque ligne de texte correspond à un noeud ou à une feuille, et la marge indique le niveau de l'arbre. Pour un noeud, la condition de division est affichée; pour une feuille, l'étiquette de classe attribuée apparaît.

### ***Pondérations d'instance et pondérations de classe***

Par défaut, il est supposé que tous les enregistrements et classes d'entrée sont de même importance relative. Vous pouvez modifier cette option en attribuant des pondérations individuelles aux membres d'un de ces éléments ou des deux. Par exemple, cela peut être utile si les points de données de vos données d'apprentissage ne sont pas distribués de manière réaliste entre les catégories. Les pondérations permettent d'orienter le modèle afin de compenser les catégories les moins représentées dans les données. L'augmentation de la pondération d'une valeur cible doit augmenter le pourcentage de prédictions correctes de cette catégorie.

Dans le nœud de modélisation Arbre de décision, vous pouvez spécifier deux types de pondérations. Les **Pondérations d'instance** attribuent une pondération à chaque ligne des données d'entrée. Les pondérations sont généralement spécifiées en tant que 1.0 dans la plupart des cas, avec des valeurs plus ou moins importantes attribuées uniquement aux cas plus ou moins importants que la majorité, par exemple :

ID d'enregistrement	Cible	Pondération d'instance
1	médicamentA	1.1
2	médicamentB	1.0
3	médicamentA	1.0
4	médicamentB	0.3

Les **pondérations de classe** attribuent une pondération à chaque catégorie du champ cible, par exemple :

Classe	Pondération de classe
médicamentA	1.0
médicamentB	1.5

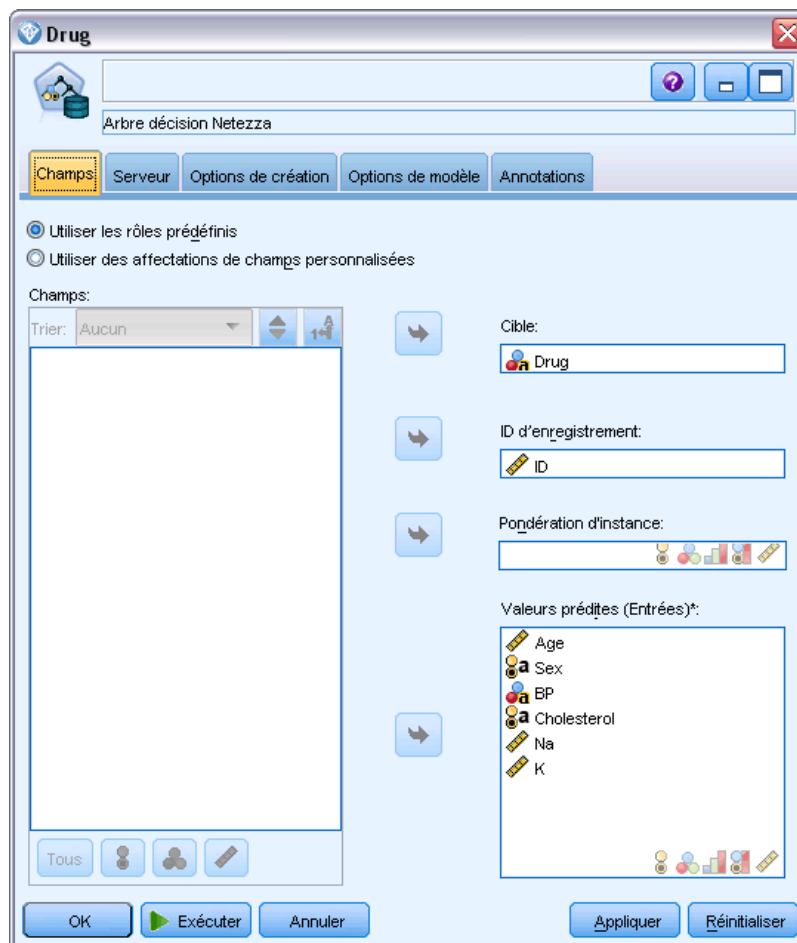
Les deux types de pondérations peuvent être utilisés en même temps, auquel cas ils sont multipliés entre eux et sont utilisés comme pondérations d'instance. Par conséquent, si les deux exemples précédents ont été utilisés ensemble, l'algorithme utilisera les pondérations d'instance suivantes.

ID d'enregistrement	Calcul	Pondération d'instance
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

## Options du champ Arbre décision Netezza

Dans l'onglet Champs, vous pouvez choisir d'utiliser les paramètres du rôle de champ déjà définis dans les noeuds en amont, ou de réaliser manuellement des affectations de champs.

Figure 6-5  
Options du champ Arbre décision



**Utiliser des rôles prédéfinis.** Cette option utilise les paramètres de rôle (cibles, variables indépendantes, etc.) pour un noeud Typier en amont (ou l'onglet Types d'un noeud source en amont). [Pour plus d'informations, reportez-vous à la section Définition du rôle du champ dans le chapitre 4 dans \*Noeuds source, exécution et de sortie de IBM SPSS Modeler 15\*.](#)

**Utiliser des affectations de champs personnalisées.** Sélectionnez cette option si vous souhaitez affecter manuellement des cibles, des variables indépendantes et d'autres rôles sur cet écran.

**Champs.** Utilisez les boutons fléchés pour affecter manuellement des éléments à partir de cette liste aux divers champs de rôle à droite de l'écran. Les icônes indiquent les niveaux de mesure valides pour chaque champ de rôle.

Cliquez sur le bouton Tous pour sélectionner tous les champs de la liste, ou cliquez sur un bouton de niveau de mesure individuelle pour sélectionner tous les champs avec un niveau de mesure.

**Cible.** Sélectionnez un champ comme cible pour la prédiction.

**ID d'enregistrement.** Le champ est utilisé comme identificateur d'enregistrement unique. Les valeurs de ce champ doivent être uniques pour chaque enregistrement (par exemple, les numéros d'ID de client).

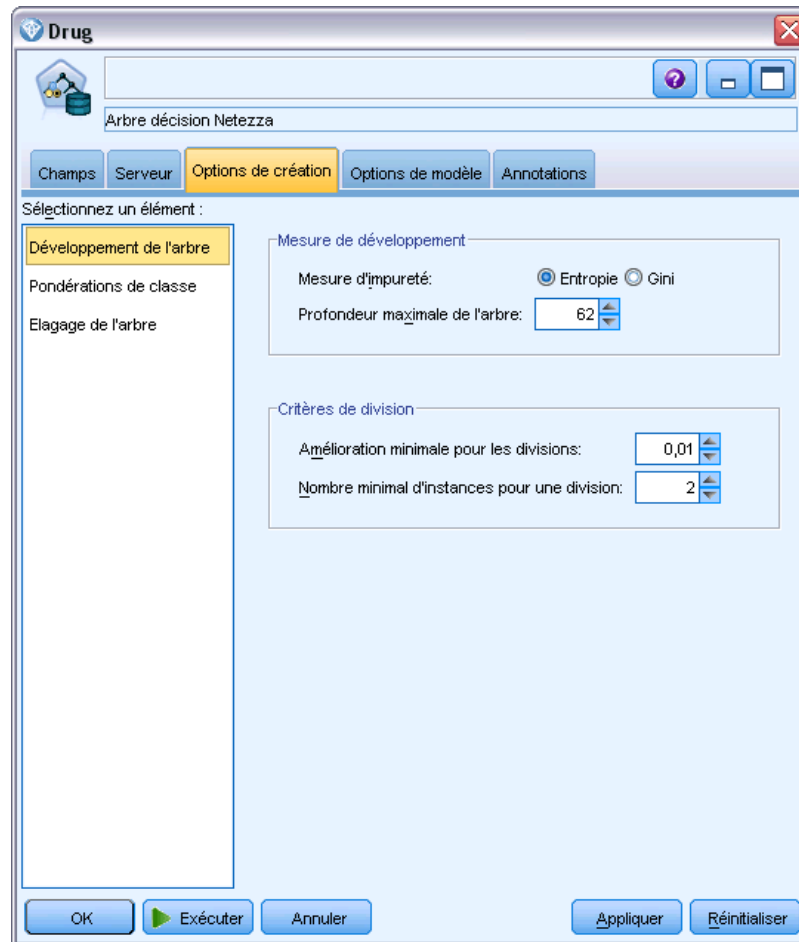
**Pondération d'instance.** Spécifier un champ ici vous permet d'utiliser les pondérations d'instance (une pondération par ligne de données d'entrée) à la place de, ou en plus, des pondérations de classe par défaut (une pondération par catégorie pour le champ cible). Le champ spécifié ici doit être un champ contenant une pondération numérique pour chaque ligne de données d'entrée. [Pour plus d'informations, reportez-vous à la section Pondérations d'instance et pondérations de classe sur p. 176.](#)

**Valeurs prédites (Entrées).** Sélectionnez le ou les champs d'entrée. Cela revient à définir le rôle du champ sur la valeur *Entrée* dans un nœud *Typar*.

### ***Options de création d'arbre décision Netezza***

L'onglet Options de création est celui où vous définissez toutes les options de création d'un modèle. Vous pouvez bien sûr cliquer simplement sur le bouton Exécuter pour créer un modèle avec toutes les options par défaut, mais normalement vous pouvez personnaliser la création selon vos propres objectifs.

Figure 6-6  
Options de création de l'arbre de décision pour le développement de l'arbre



Vous pouvez définir des options de création pour :

- Développement de l'arbre.
- Pondérations pour les étiquettes de classe
- Elagage de l'arbre

Les options pour le développement de l'arbre sont décrites dans cette section.

**Mesure de développement** . Ces options contrôlent la manière dont le développement de l'arbre est mesuré. Si vous ne voulez pas utiliser les valeurs par défaut, cliquez sur Personnaliser et effectuez les changements.

- **Mesure d'impureté.** La mesure d'impureté, utilisée pour évaluer le meilleur endroit où diviser l'arbre. L'**impureté** correspond à la mesure de la variabilité, dans les sous-groupes définis par l'arbre, des valeurs de leurs champs de sortie.

Les valeurs numériques prises en charge sont **Entropy** (défaut) et **Gini**. Ce sont deux mesures d'impureté connues dont le calcul est basé sur les probabilités d'appartenance à une catégorie de la branche.

- **Profondeur maximale de l'arbre.** Le nombre maximal de niveaux que l'arbre peut atteindre en se développant au-dessous du nœud racine (ce qui revient au nombre de fois où l'échantillon va être divisé de manière récursive). La valeur par défaut est 62, ce qui est la profondeur d'arbre maximale à des fins de modélisation. Remarquez toutefois que le visualiseur dans le nugget de modèle peut afficher un maximum de 10 niveaux.

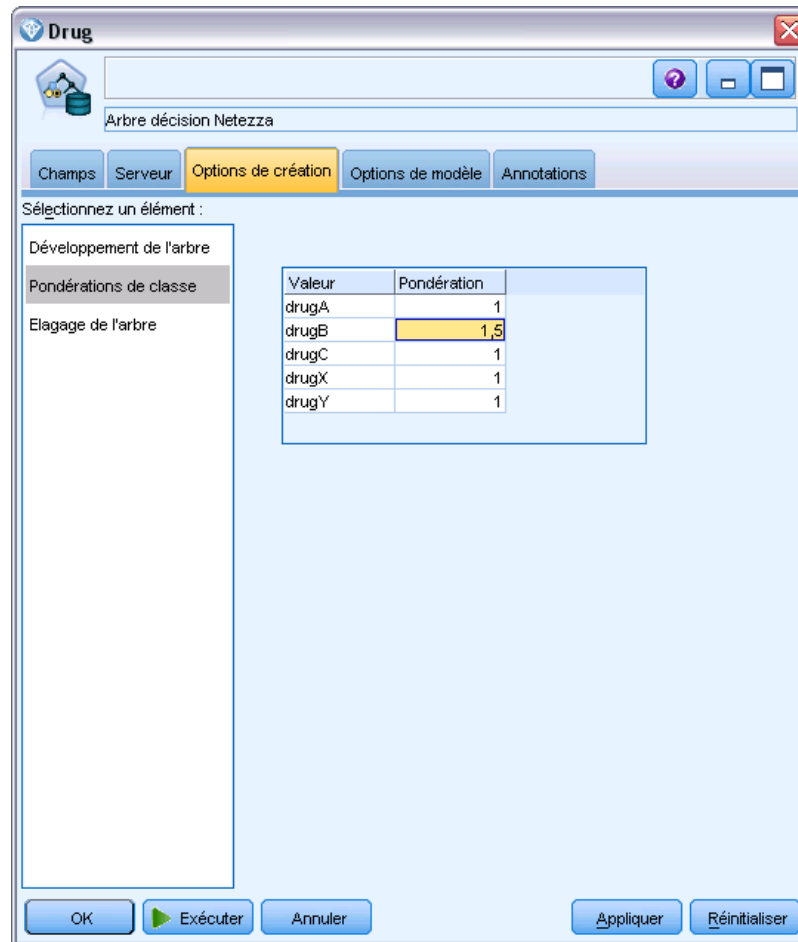
**Critères de division.** Ces options permettent de contrôler l'arrêt de la division de l'arbre. Si vous ne voulez pas utiliser les valeurs par défaut, cliquez sur Personnaliser et effectuez les changements.

- **Amélioration minimale pour les divisions.** La quantité minimale d'impureté devant être réduite avant qu'une nouvelle division puisse être créée dans l'arbre. La création d'arbre vise à créer des sous-groupes ayant des résultats similaires, c'est-à-dire à minimiser l'impureté de chaque nœud. Si le meilleur découpage calculé pour une branche signifie une réduction de l'impureté inférieure à la valeur spécifiée par le critère de division, alors le découpage ne sera pas effectué.
- **Nombre minimal d'instances pour une division.** Le nombre minimal d'enregistrements pouvant être divisés. Lorsqu'il reste moins d'enregistrements que ce nombre d'enregistrements non divisés, aucune division supplémentaire ne sera effectuée. Ce champ peut être utilisé pour éviter la création de sous-groupes minuscules dans l'arbre.

### ***Nœud Arbre décision Netezza - Pondérations de classe***

Vous pouvez assigner ici des pondérations à des classes individuelles. Par défaut une valeur de 1 est attribuée à toutes les classes, pour les rendre pondérées de manière identique. En spécifiant des pondérations numériques différentes pour des étiquettes de classes différentes, les ensembles d'apprentissage des classes spécifiques seront pondérés en conséquence par l'algorithme.

Figure 6-7  
Options de pondération de classe de l'arbre décision



Pour modifier une pondération, double-cliquez dessus dans la colonne Pondération et effectuez les modifications désirées.

**Valeur.** L'ensemble d'étiquettes de classe calculé à partir des valeurs possibles du champ cible.

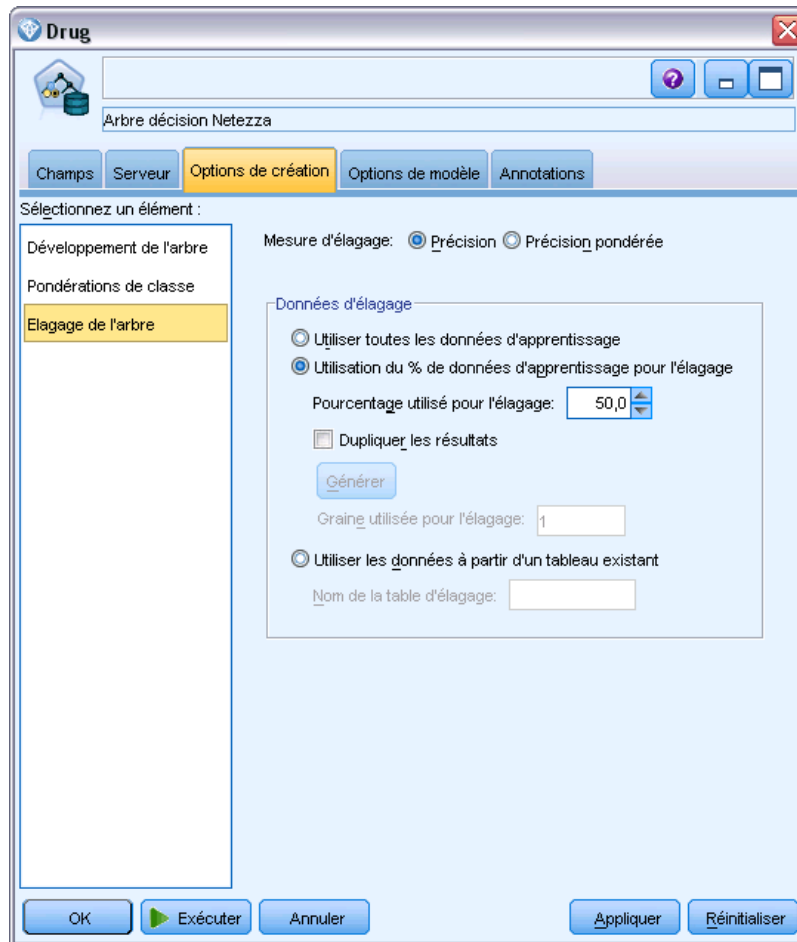
**Pondération.** La pondération à attribuer à une classe spécifique. Affecter une pondération supérieure à une classe rend le modèle plus sensible à cette classe par rapport aux autres classes.

Vous pouvez utiliser les pondérations de classe en combinaison avec les pondérations d'instance. [Pour plus d'informations, reportez-vous à la section Pondérations d'instance et pondérations de classe sur p. 176.](#)

### ***Nœud Arbre décision Netezza - Élagage de l'arbre***

Vous pouvez utiliser les options d'élagage pour spécifier les critères d'élagage de l'arbre décision. Le but de l'élagage est de réduire le risque de surajustement en supprimant les sous-groupes trop développés qui n'améliorent pas la précision attendue des nouvelles données.

Figure 6-8  
Options d'élagage de l'arbre décision



**Mesure d'élagage.** La valeur par défaut de la mesure d'élagage, Précision, garantit que la précision du modèle demeure dans des limites acceptables après avoir supprimé une feuille de l'arbre. Utilisez l'alternative, Précision pondérée, si vous désirez prendre en compte les pondérations des classes lorsque vous effectuez l'élagage.

**Données pour l'élagage.** Vous pouvez utiliser une partie ou la totalité des données d'apprentissage pour estimer la précision attendue des nouvelles données. Vous pouvez également utiliser un ensemble de données d'élagage distinct d'une table spécifique à cette fin.

- **Utiliser toutes les données d'apprentissage.** Cette option (option par défaut) utilise toutes les données d'apprentissage pour estimer la précision du modèle.
- **Utiliser % des données d'apprentissage pour l'élagage.** Utilisez cette option pour diviser les données en deux ensembles, un pour l'apprentissage et un pour l'élagage, à l'aide du pourcentage spécifié ici pour les données d'élagage.



Sélectionnez Dupliquer les résultats si vous voulez spécifier une graine aléatoire pour vous assurer que les données sont partitionnées de la même façon chaque fois que vous exécutez le flux. Vous pouvez spécifier un entier dans le champ Graine utilisée pour l'élagage, ou bien cliquer sur Générer, ce qui créera un entier pseudo-aléatoire.

- **Utiliser des données d'une table existante.** Spécifiez le nom de la table d'un ensemble de données d'élagage distinct pour estimer la précision du modèle. Ceci est plus fiable que d'utiliser des données d'apprentissage. Toutefois, cette option peut provoquer la suppression d'un vaste sous-ensemble de données de l'ensemble d'apprentissage, ce qui réduit la qualité de l'arbre décision.

## **Netezza K-means**

Le nœud K-Means implémente l'algorithme *k*-means qui propose une méthode d'analyse de classes. Vous pouvez utiliser ce nœud pour classer un ensemble de données en groupes distincts.

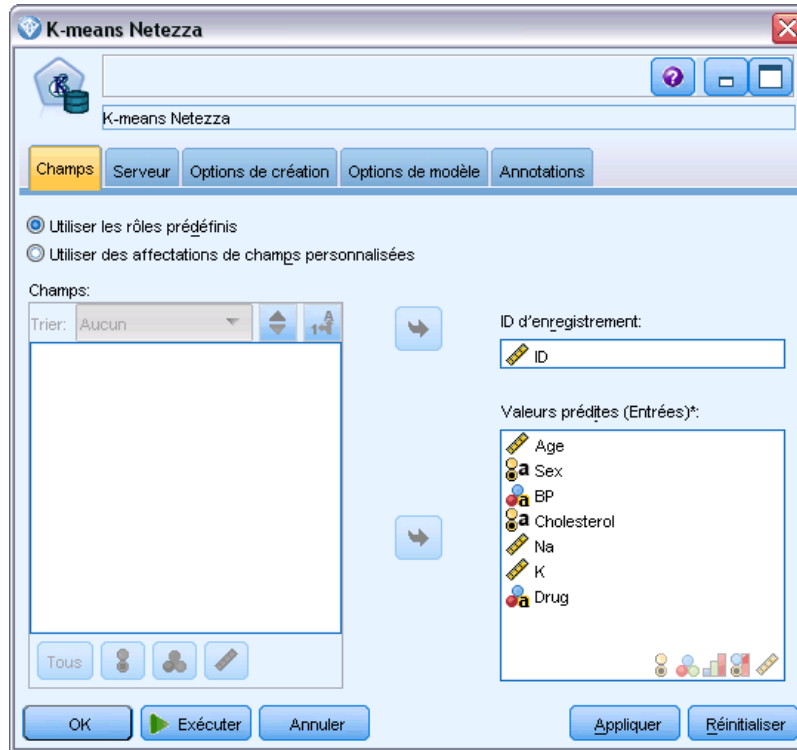
Cet algorithme est un algorithme de classification basé sur les distances qui utilise une (fonction de) mesure des distances permettant de mesurer la similarité entre les points de données. Les points de données sont affectés à la classe la plus proche, en fonction de la mesure de distance utilisée.

Cet algorithme fonctionne en effectuant plusieurs itérations du même processus de base, dans lequel chaque instance d'apprentissage est attribuée à la classe la plus proche (en respectant la fonction de distance spécifiée, appliquée au centre de l'instance et de la classe). Tous les centres de classe sont alors recalculés en tant que vecteurs de valeur d'attribut moyenne des instances attribuées à des classes spécifiques.

### **Options du champ K-Means Netezza**

Dans l'onglet Champs, vous pouvez choisir d'utiliser les paramètres du rôle de champ déjà définis dans les nœuds en amont, ou de réaliser manuellement des affectations de champs.

Figure 6-9  
Options du champ K-Means



**Utiliser des rôles prédéfinis.** Cette option utilise les paramètres de rôle (cibles, variables indépendantes, etc.) pour un noeud Typier en amont (ou l'onglet Types d'un noeud source en amont). [Pour plus d'informations, reportez-vous à la section Définition du rôle du champ dans le chapitre 4 dans \*Noeuds source, exécution et de sortie de IBM SPSS Modeler 15\*.](#)

**Utiliser des affectations de champs personnalisées.** Sélectionnez cette option si vous souhaitez affecter manuellement des cibles, des variables indépendantes et d'autres rôles sur cet écran.

**Champs.** Utilisez les boutons fléchés pour affecter manuellement des éléments à partir de cette liste aux divers champs de rôle à droite de l'écran. Les icônes indiquent les niveaux de mesure valides pour chaque champ de rôle.

Cliquez sur le bouton Tous pour sélectionner tous les champs de la liste ou cliquez sur un bouton de niveau de mesure individuelle pour sélectionner tous les champs avec un niveau de mesure.

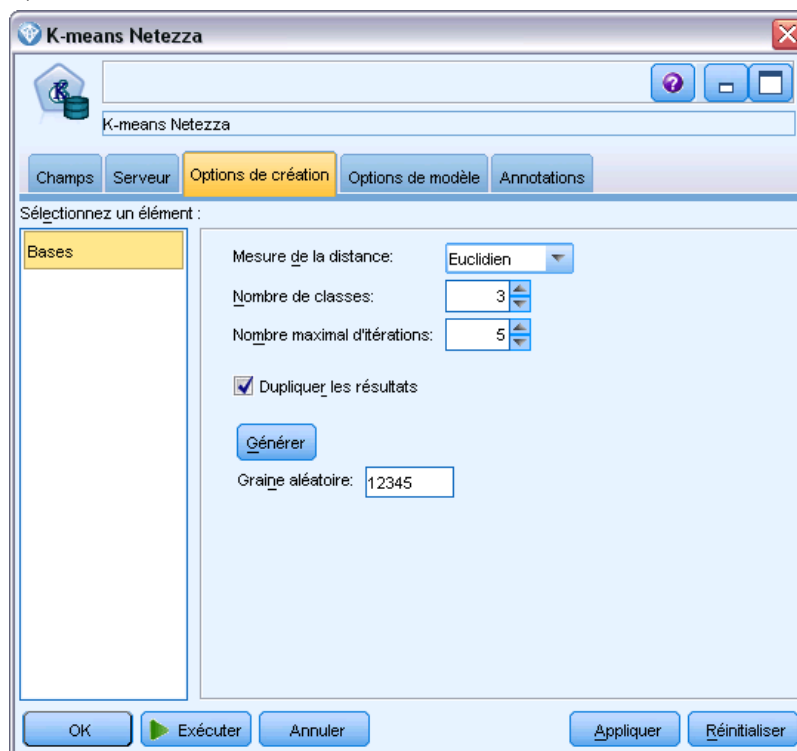
**ID d'enregistrement.** Le champ est utilisé comme identificateur d'enregistrement unique.

**Variables indépendantes (Entrées).** Sélectionnez un ou plusieurs champs comme entrée pour la prédiction.

## Options de création K-Means Netezza

L'onglet Options de création est celui où vous définissez toutes les options de création d'un modèle. Vous pouvez bien sûr cliquer simplement sur le bouton Exécuter pour créer un modèle avec toutes les options par défaut, mais normalement vous pouvez personnaliser la création selon vos propres objectifs.

Figure 6-10  
Options de création K-Means



**Mesure de distance.** La méthode à utiliser pour la mesure de distance entre les points de données ; plus les distances sont importantes plus les différences le sont aussi. Les options sont les suivantes :

- **Euclidienne.** (par défaut) La distance entre deux points est calculée en les reliant par une ligne droite.
- **Manhattan.** La distance entre deux points est calculée en tant que total des différences absolues entre leur coordonnées.
- **Canberra.** Semblable à la distance Manhattan, mais plus sensible aux points de données plus proches de l'origine.
- **Maximum :** la distance entre deux points est calculée comme la plus importante de leurs différences parmi toutes les dimensions de leurs coordonnées.

**Nombre de classes (k).** Indiquez le nombre de classes à créer.

**Nombre maximum des itérations.** Cet algorithme fonctionne en effectuant plusieurs itérations du même processus. Cette option permet d'interrompre l'apprentissage du modèle lorsque le nombre d'itérations spécifié est atteint.

**Dupliquer les résultats.** Cochez cette case si vous souhaitez définir une graine aléatoire qui vous permettra de dupliquer les analyses. Vous pouvez spécifier un entier ou cliquer sur Générer, ce qui crée un entier pseudo-aléatoire.

## ***Bayes Net Netezza***

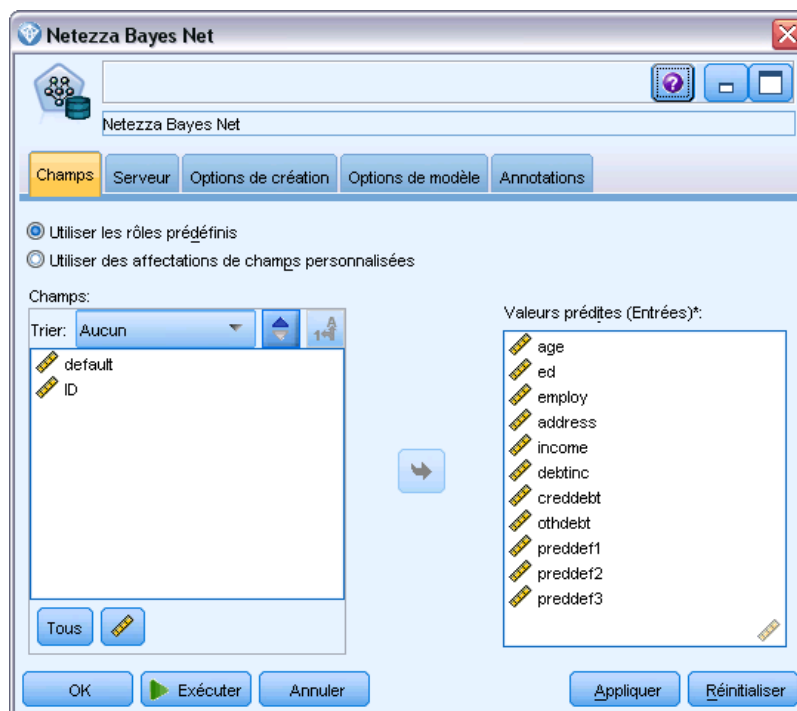
Un réseau bayésien est un modèle qui affiche les variables dans un ensemble de données et les indépendances probabilistes ou conditionnelles entre elles. Avec le nœud Bayes Net Netezza, vous pouvez créer un modèle de probabilité en combinant les preuves observées et enregistrées avec les connaissances réelles « de bon sens » pour établir la probabilité des occurrences en utilisant des attributs apparemment sans lien.

### ***Options de champs du réseau Bayes Net Netezza***

Dans l'onglet Champs, vous pouvez choisir d'utiliser les paramètres du rôle de champ déjà définis dans les nœuds en amont, ou de réaliser manuellement des affectations de champs.

Pour ce nœud, le champ cible est uniquement requis pour le scoring, il n'est donc pas affiché dans cet onglet. Vous pouvez définir ou modifier la cible sur un nœud Typier, dans l'onglet Options des modèles de ce nœud, ou dans l'onglet Paramètres du nugget de modèle. [Pour plus d'informations, reportez-vous à la section Nugget Bayes Net Netezza - Onglet Paramètres sur p. 226.](#)

Figure 6-11  
Options de champs Bayes Net



**Utiliser des rôles prédéfinis.** Cette option utilise les paramètres de rôle (cibles, variables indépendantes, etc.) pour un nœud Typer en amont (ou l'onglet Types d'un nœud source en amont). Pour plus d'informations, reportez-vous à la section *Définition du rôle du champ* dans le chapitre 4 dans *Noeuds source, exécution et de sortie de IBM SPSS Modeler 15*.

**Utiliser des affectations de champs personnalisées.** Sélectionnez cette option si vous souhaitez affecter manuellement des cibles, des variables indépendantes et d'autres rôles sur cet écran.

**Champs.** Utilisez les boutons fléchés pour affecter manuellement des éléments à partir de cette liste aux divers champs de rôle à droite de l'écran. Les icônes indiquent les niveaux de mesure valides pour chaque champ de rôle.

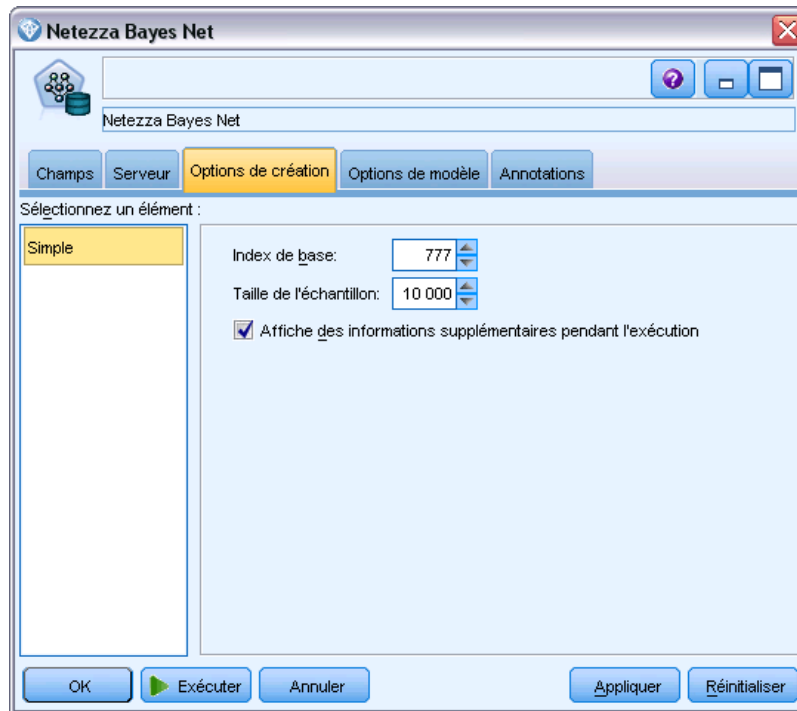
Cliquez sur le bouton Tous pour sélectionner tous les champs de la liste ou cliquez sur un bouton de niveau de mesure individuelle pour sélectionner tous les champs avec un niveau de mesure.

**Valeurs prédites (Entrées).** Sélectionnez un ou plusieurs champs comme entrée pour la prédiction.

### ***Options de création du réseau Bayes Net Netezza***

L'onglet Options de création est celui où vous définissez toutes les options de création d'un modèle. Vous pouvez bien sûr cliquer simplement sur le bouton Exécuter pour créer un modèle avec toutes les options par défaut, mais normalement vous pouvez personnaliser la création selon vos propres objectifs.

Figure 6-12  
Options de création Bayes Net



**Indice de base :** l'identifiant numérique à affecter au premier attribut (champ d'entrée) pour une gestion interne simplifiée.

**Taille de l'échantillon :** la taille de l'échantillon à prélever si le nombre d'attributs est si grand que le temps de traitement serait inacceptablement long.

**Afficher des informations supplémentaires au cours de l'exécution :** si cette case est cochée (par défaut), des informations de progression supplémentaires s'affichent dans une boîte de dialogue.

## Naive Bayes Netezza

Naive Bayes est un algorithme bien connu pour les problèmes de classification. Ce modèle est appelé *naïve*, car il considère toutes les variables de prévision proposées comme étant indépendantes les unes des autres. Naive Bayes est un algorithme rapide et évolutif qui calcule les probabilités conditionnelles des combinaisons d'attributs et de l'attribut cible. Une probabilité indépendante est établie à partir des données d'apprentissage. Cette probabilité détermine la vraisemblance de chaque classe cible et est calculée en fonction de l'occurrence de chaque catégorie de valeur issue des variables d'entrée.

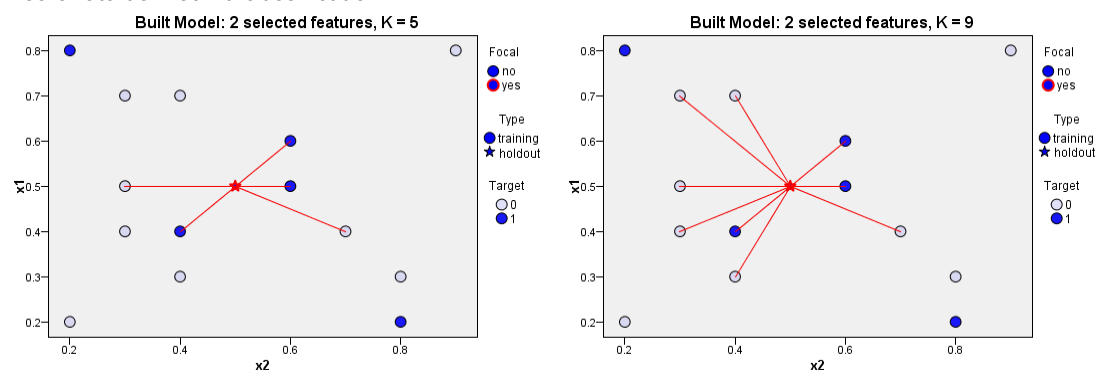
## KNN Netezza

L'analyse d'agrégation suivant le saut minimum (ou du plus proche voisin) est une méthode de classification des observations basée sur la similarité des observations entre elles. Dans le domaine de l'apprentissage automatique, elle a été développée comme un moyen de reconnaître des patrons de données sans nécessiter une correspondance exacte à une observation ou à un patron enregistré. Les observations semblables sont proches les unes des autres et les observations dissemblables sont éloignées les unes des autres. Ainsi la distance entre deux observations est une mesure de leur dissimilarité.

Les observations proches les unes des autres sont appelées « voisins ». Lorsqu'une nouvelle observation (de rétention) est présentée, sa distance de chaque observation du modèle est calculée. Les classifications des observations les plus similaires « les plus proches voisins » sont mesurées et la nouvelle observation est placée dans la catégorie qui contient le plus grand nombre de voisins les plus proches.

Vous pouvez spécifier le nombre de voisins les plus proches à examiner, cette valeur est appelée  $k$ . Les images montrent comment une nouvelle observation est classifiée en utilisant deux valeurs différentes de  $k$ . Lorsque  $k = 5$ , la nouvelle observation est placée dans la catégorie 1 car une majorité de voisins les plus proches appartiennent à la catégorie 1. Toutefois, lorsque  $k = 9$ , la nouvelle observation est placée dans la catégorie 0 car une majorité de voisins les plus proches appartiennent à la catégorie 0.

Figure 6-13  
Les effets de  $k$  sur la classification

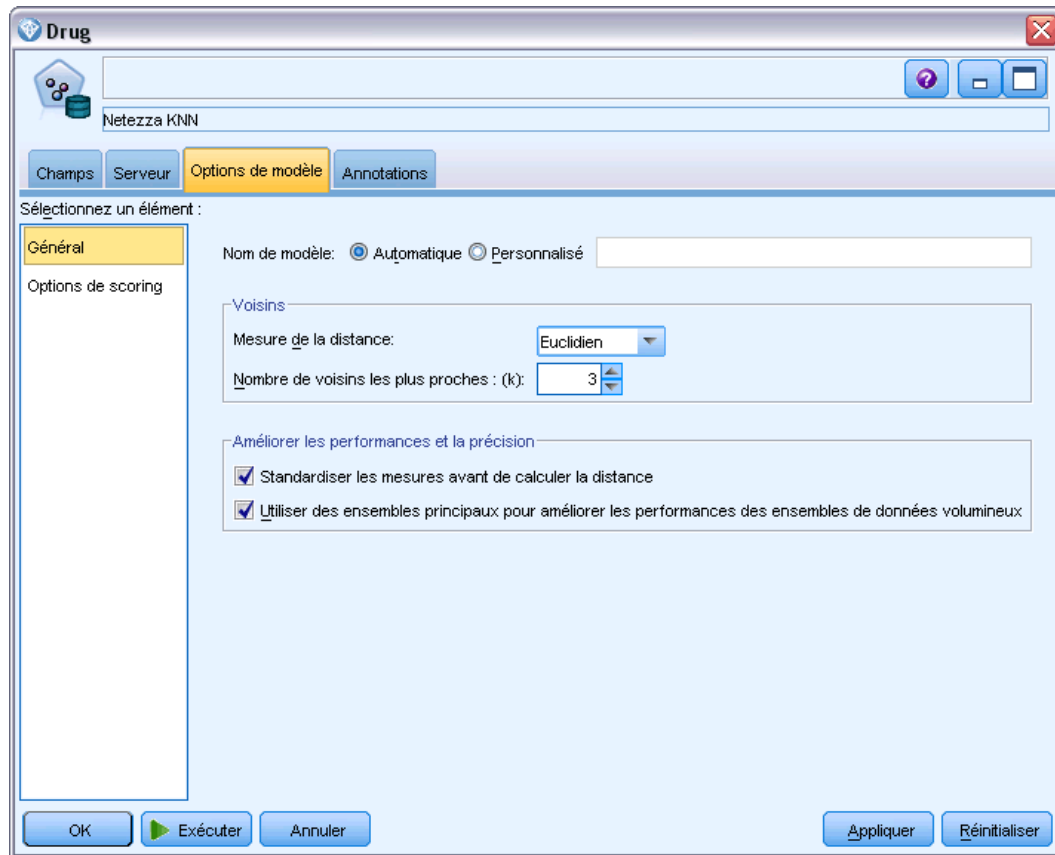


L'analyse d'agrégation suivant le saut minimum peut aussi être utilisée pour calculer les valeurs d'une cible continue. Dans cette situation, la valeur cible moyenne ou médiane des voisins les plus proches est utilisée pour obtenir la valeur prédite de la nouvelle observation.

## Options de modèle KNN Netezza - Général

Dans l'onglet Options de modèle - Général, vous pouvez choisir de spécifier un nom pour le modèle ou de générer automatiquement un nom. Vous pouvez aussi définir des options qui contrôlent la manière de calculer le nombre d'agrégations suivant le saut minimum et des options pour une meilleure performance et une meilleure précision du modèle.

Figure 6-14  
Options générales de modèle KNN



**Nom du modèle.** Vous pouvez générer le nom du modèle automatiquement sur la base du champ cible ou ID (ou du type de modèle si aucun de ces champs n'est spécifié) ou spécifier un nom personnalisé.

### Voisins

**Mesure de distance.** La méthode à utiliser pour la mesure de distance entre les points de données ; plus les distances sont importantes plus les différences le sont aussi. Les options sont les suivantes :

- **Euclidienne.** (par défaut) La distance entre deux points est calculée en les reliant par une ligne droite.
- **Manhattan.** La distance entre deux points est calculée en tant que total des différences absolues entre leur coordonnées.
- **Canberra.** Semblable à la distance Manhattan, mais plus sensible aux points de données plus proches de l'origine.
- **Maximum :** la distance entre deux points est calculée comme la plus importante de leurs différences parmi toutes les dimensions de leurs coordonnées.



**Nombre de plus proches voisins (k).** Le nombre de plus proches voisins pour une observation particulière. Remarque : l'utilisation d'un nombre élevé de voisins ne garantit pas forcément un modèle plus précis.

Le choix de  $k$  contrôle l'équilibre entre la prévention du surajustement (ceci peut être important, en particulier pour les données parasites) et la résolution (donnant différentes prévisions pour des instances semblables). Vous devrez généralement régler la valeur de  $k$  pour chaque ensemble de données, avec des valeurs typiques allant de 1 à plusieurs douzaines.

#### ***Améliorer la performance et la précision***

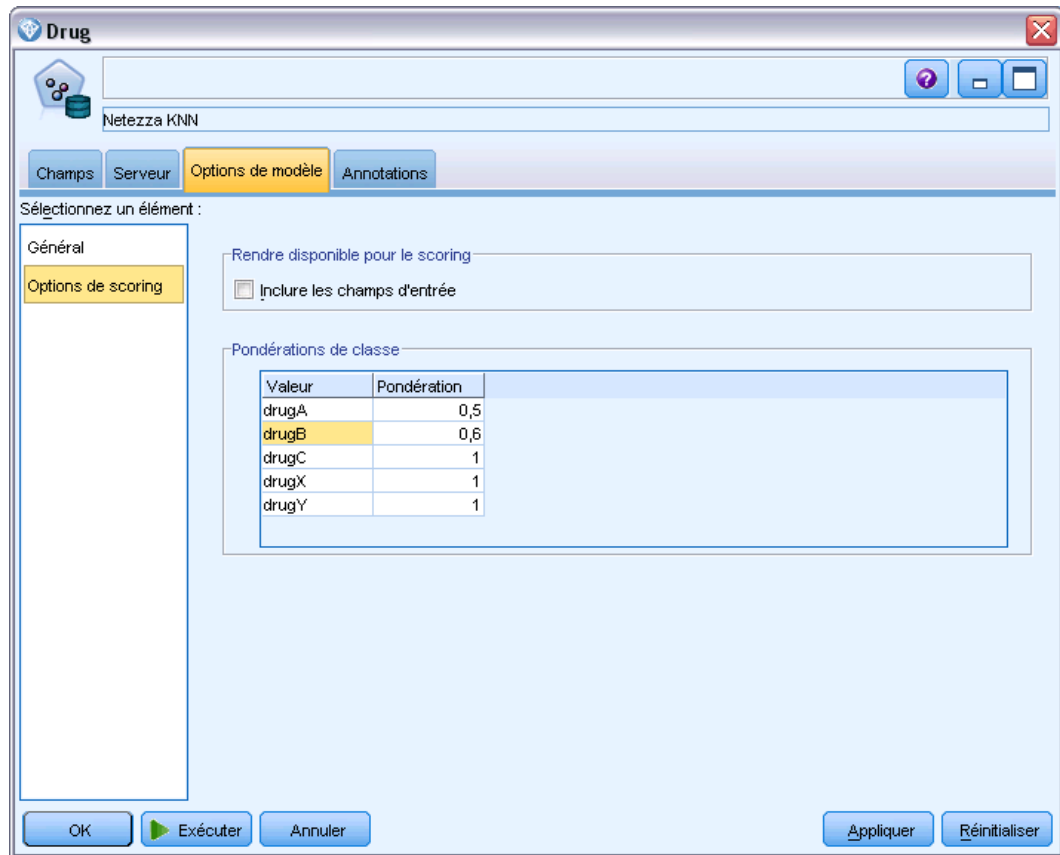
**Standardisez les mesures avant de calculer la distance.** Si elle est sélectionnée, cette option standardise les mesures des champs d'entrées continues avant de calculer les valeurs de distance.

**Utilisez des ensembles principaux pour améliorer les performances pour de grands ensembles de données.** Si elle est sélectionnée, cette option utilise l'échantillonnage d'ensembles principaux pour accélérer le calcul quand de grands ensembles de données sont impliqués.

### ***Options de modèle KNN Netezza - Options de scoring***

Dans l'onglet Options de modèle - Options de scoring, vous pouvez définir la valeur par défaut pour une option de scoring et attribuer des pondérations relatives aux classes individuelles.

Figure 6-15  
Options générales de modèle KNN



### ***Rendre disponible pour le scoring***

**Inclure les champs d'entrée.** Spécifie si les champs d'entrée sont inclus dans le scoring par défaut.

### ***Pondérations de classe***

Utilisez cette option pour modifier l'importance relative de chaque classe lors de la création du modèle.

*Remarque* : cette option est uniquement activée si vous utilisez KNN pour la classification. Si vous effectuez une régression (si le type de champ cible est Continu), l'option est désactivée.

Par défaut une valeur de 1 est attribuée à toutes les classes, pour les rendre pondérées de manière identique. En spécifiant des pondérations numériques différentes pour des étiquettes de classes différentes, les ensembles d'apprentissage des classes spécifiques seront pondérés en conséquence par l'algorithme.

Pour modifier une pondération, double-cliquez dessus dans la colonne Pondération et effectuez les modifications désirées.

**Valeur.** L'ensemble d'étiquettes de classe calculé à partir des valeurs possibles du champ cible.

**Pondération.** La pondération à attribuer à une classe spécifique. Affecter une pondération supérieure à une classe rend le modèle plus sensible à cette classe par rapport aux autres classes.

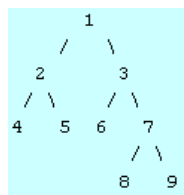
## Classification par division Netezza

La classification par division est une méthode d'analyse des classes dans laquelle l'algorithme est exécuté de manière répétitive pour diviser les classes en sous-classes jusqu'à atteindre un point d'arrêt spécifié.

La formation de classes commence par une seule classe contenant toutes les instances de formation (enregistrements). La première itération de l'algorithme divise l'ensemble de données en deux sous-classes, avec des itérations consécutives qui les divisent encore en sous-classes. Les critères d'arrêt sont spécifiés sous la forme d'un nombre maximum d'itérations, d'un nombre maximum de niveaux dans lesquels l'ensemble de données est divisé et d'un nombre minimum requis d'instances pour une partition supplémentaire.

L'arbre de classification hiérarchique résultant peut être utilisé pour classifier les instances en les propageant à partir de la classe racine, comme dans l'exemple suivant.

Figure 6-16  
Exemple d'arbre de classification par division



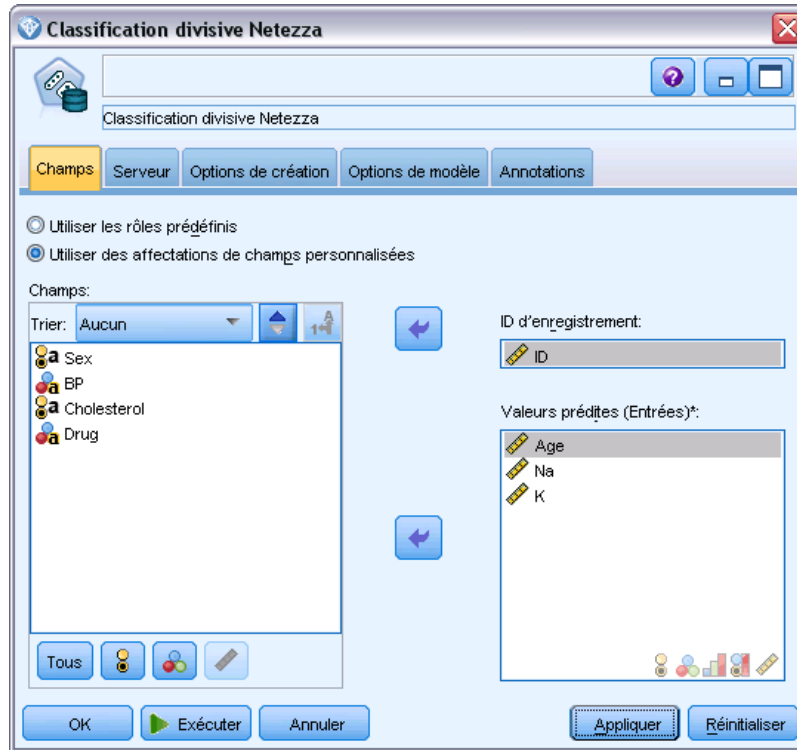
À chaque niveau, la meilleure sous-classe correspondante est choisie par rapport à la distance entre l'instance et les centres de la sous-classe.

Quand les instances sont évaluées par un niveau hiérarchique appliqué de -1 (par défaut), le scoring renvoie uniquement une classe de feuilles, car les feuilles sont désignées par un nombre négatif. Dans l'exemple, il s'agirait des classes 4, 5, 6, 8 ou 9. Cependant, si le niveau hiérarchique est par exemple défini à 2, le scoring renvoie l'une des classes au deuxième niveau en dessous de la classe racine, c'est-à-dire 4, 5, 6 ou 7.

## Options de champ de classification par division Netezza

Dans l'onglet Champs, vous pouvez choisir d'utiliser les paramètres du rôle de champ déjà définis dans les noeuds en amont, ou de réaliser manuellement des affectations de champs.

Figure 6-17  
Options de champ Classification par division



**Utiliser des rôles prédéfinis.** Cette option utilise les paramètres de rôle (cibles, variables indépendantes, etc.) pour un noeud Typier en amont (ou l'onglet Types d'un noeud source en amont). [Pour plus d'informations, reportez-vous à la section Définition du rôle du champ dans le chapitre 4 dans \*Noeuds source, exécution et de sortie de IBM SPSS Modeler 15\*.](#)

**Utiliser des affectations de champs personnalisées.** Sélectionnez cette option si vous souhaitez affecter manuellement des cibles, des variables indépendantes et d'autres rôles sur cet écran.

**Champs.** Utilisez les boutons fléchés pour affecter manuellement des éléments à partir de cette liste aux divers champs de rôle à droite de l'écran. Les icônes indiquent les niveaux de mesure valides pour chaque champ de rôle.

Cliquez sur le bouton Tous pour sélectionner tous les champs de la liste ou cliquez sur un bouton de niveau de mesure individuelle pour sélectionner tous les champs avec un niveau de mesure.

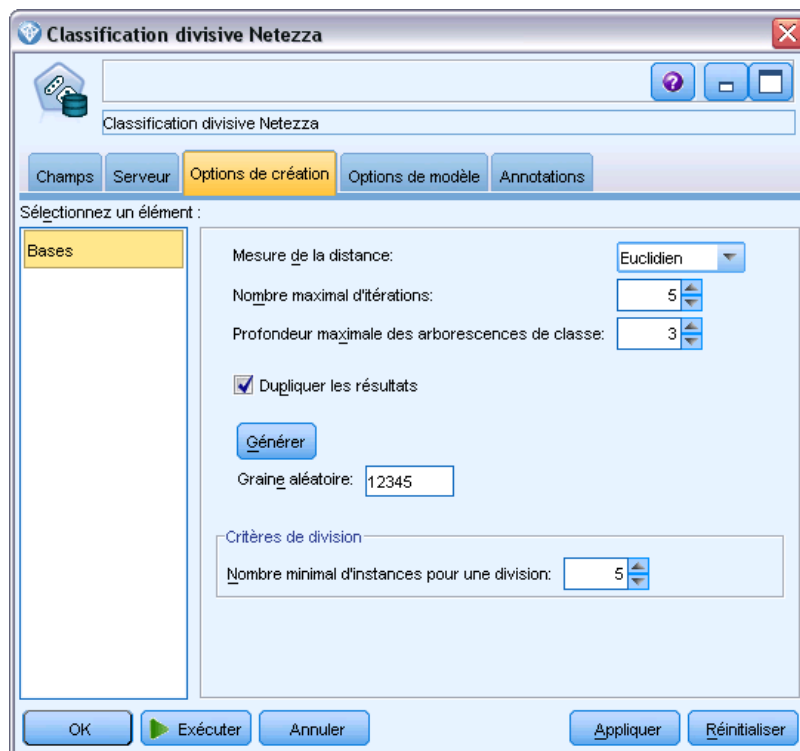
**ID d'enregistrement.** Le champ est utilisé comme identificateur d'enregistrement unique.

**Variables indépendantes (Entrées).** Sélectionnez un ou plusieurs champs comme entrée pour la prédiction.

## Options de création de classification par division Netezza

L'onglet Options de création est celui où vous définissez toutes les options de création d'un modèle. Vous pouvez bien sûr cliquer simplement sur le bouton Exécuter pour créer un modèle avec toutes les options par défaut, mais normalement vous pouvez personnaliser la création selon vos propres objectifs.

Figure 6-18  
Options de création Classification par division



**Mesure de distance.** La méthode à utiliser pour la mesure de distance entre les points de données ; plus les distances sont importantes plus les différences le sont aussi. Les options sont les suivantes :

- **Euclidienne.** (par défaut) La distance entre deux points est calculée en les reliant par une ligne droite.
- **Manhattan.** La distance entre deux points est calculée en tant que total des différences absolues entre leur coordonnées.
- **Canberra.** Semblable à la distance Manhattan, mais plus sensible aux points de données plus proches de l'origine.
- **Maximum :** la distance entre deux points est calculée comme la plus importante de leurs différences parmi toutes les dimensions de leurs coordonnées.

**Nombre maximum des itérations.** Cet algorithme fonctionne en effectuant plusieurs itérations du même processus. Cette option permet d'interrompre l'apprentissage du modèle lorsque le nombre d'itérations spécifié est atteint.

**Profondeur maximale des arbres de classification.** Le nombre maximal de niveaux dans lesquels l'ensemble de données peut être subdivisé.

**Dupliquer les résultats.** Cochez cette case si vous souhaitez définir une graine aléatoire qui vous permettra de dupliquer les analyses. Vous pouvez spécifier un entier ou cliquer sur Générer, ce qui crée un entier pseudo-aléatoire.

**Nombre minimal d'instances pour une division.** Le nombre minimal d'enregistrements pouvant être divisés. Lorsqu'il reste moins d'enregistrements que ce nombre d'enregistrements non divisés, aucune division supplémentaire ne sera effectuée. Ce champ peut être utilisé pour éviter la création de sous-groupes minuscules dans l'arbre de classification.

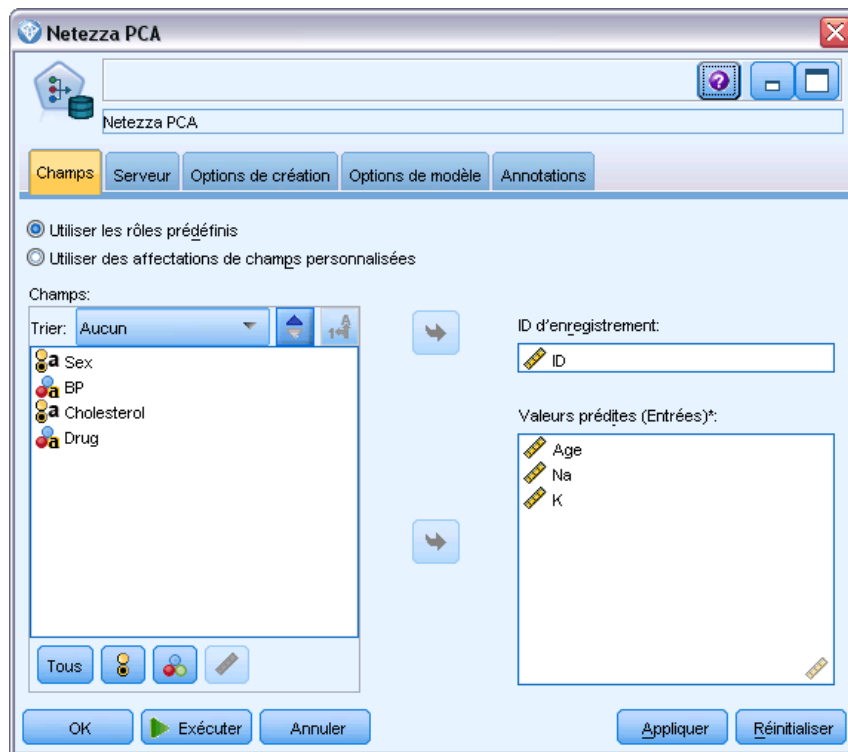
## ***ACP Netezza***

L'analyse en composantes principales (ACP) est une technique de réduction des données performante conçue pour réduire la complexité des données. L'ACP recherche les combinaisons linéaires des champs d'entrée qui permettent de capturer au mieux la variance dans l'ensemble des champs, où les composantes sont orthogonales (non corrélées) les unes aux autres. Le but consiste à trouver un nombre limité de champs dérivés (les composantes principales) récapitulant les informations contenues dans l'ensemble de champs d'origine.

### ***Options de champs ACP Netezza***

Dans l'onglet Champs, vous pouvez choisir d'utiliser les paramètres du rôle de champ déjà définis dans les noeuds en amont, ou de réaliser manuellement des affectations de champs.

Figure 6-19  
Options de champs ACP



**Utiliser des rôles prédéfinis.** Cette option utilise les paramètres de rôle (cibles, variables indépendantes, etc.) pour un noeud Typier en amont (ou l'onglet Types d'un noeud source en amont). [Pour plus d'informations, reportez-vous à la section Définition du rôle du champ dans le chapitre 4 dans \*Noeuds source, exécution et de sortie de IBM SPSS Modeler 15\*.](#)

**Utiliser des affectations de champs personnalisées.** Sélectionnez cette option si vous souhaitez affecter manuellement des cibles, des variables indépendantes et d'autres rôles sur cet écran.

**Champs.** Utilisez les boutons fléchés pour affecter manuellement des éléments à partir de cette liste aux divers champs de rôle à droite de l'écran. Les icônes indiquent les niveaux de mesure valides pour chaque champ de rôle.

Cliquez sur le bouton Tous pour sélectionner tous les champs de la liste ou cliquez sur un bouton de niveau de mesure individuelle pour sélectionner tous les champs avec un niveau de mesure.

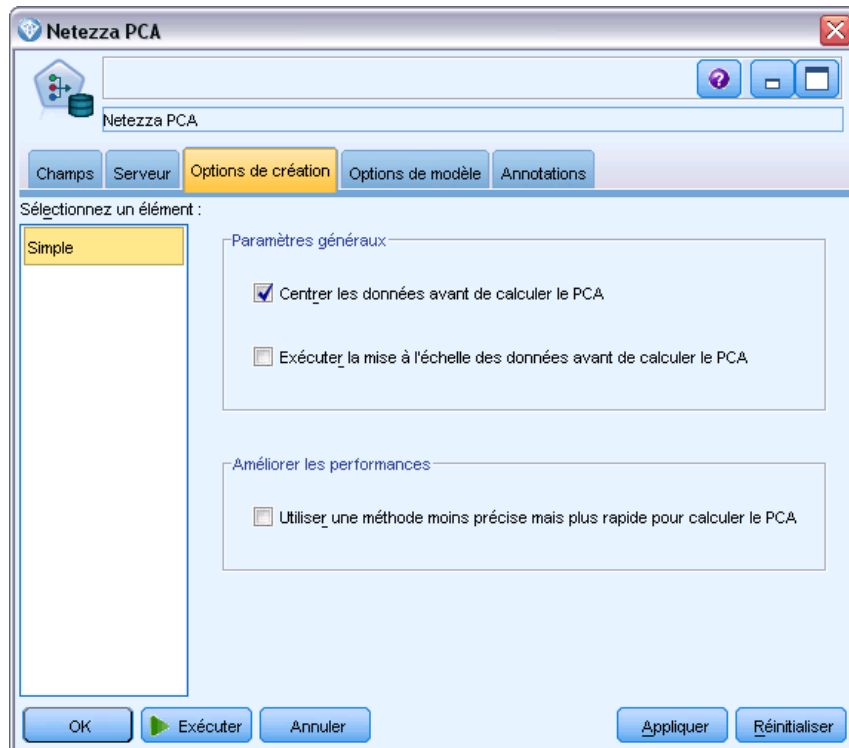
**ID d'enregistrement.** Le champ est utilisé comme identificateur d'enregistrement unique.

**Variables indépendantes (Entrées).** Sélectionnez un ou plusieurs champs comme entrée pour la prédiction.

## Options de création ACP Netezza

L'onglet Options de création est celui où vous définissez toutes les options de création d'un modèle. Vous pouvez bien sûr cliquer simplement sur le bouton Exécuter pour créer un modèle avec toutes les options par défaut, mais normalement vous pouvez personnaliser la création selon vos propres objectifs.

Figure 6-20  
Options de création ACP



**Centrer les données avant de calculer l'ACP.** Si elle est cochée (par défaut), cette option effectue un centrage des données (également nommé « soustraction moyenne ») avant l'analyse. Le centrage des données est nécessaire pour assurer que la première composante principale décrit la direction de la variance maximale, sinon la composante peut correspondre plus étroitement à la moyenne des données. Normalement, vous n'êtes censé décocher cette option que pour améliorer les performances dans le cas où les données ont déjà été préparées ainsi.

**Mettre les données à l'échelle avant de calculer l'ACP.** Cette option effectue la mise à l'échelle des données avant l'analyse. Ceci peut rendre l'analyse moins arbitraire quand différentes variables sont mesurées dans différentes unités. Dans sa forme la plus simple, la mise à l'échelle des données est possible en divisant chaque variable par sa variation standard.

**Utiliser une méthode moins précise mais plus rapide pour calculer l'ACP.** Avec cette option, l'algorithme utilise une méthode moins précise mais plus rapide (forceEigensolve) pour trouver les composantes principales.



## **Arbre de régression Netezza**

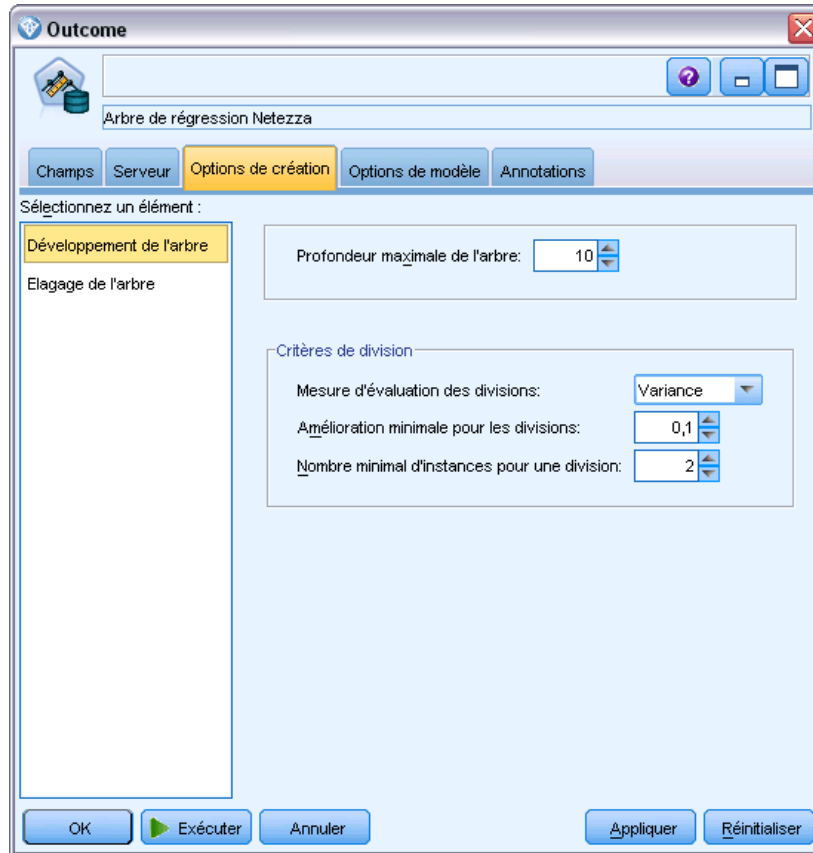
Un arbre de régression est un algorithme en arborescence qui divise l'échantillon d'observations de façon répétée afin de calculer des sous-ensembles du même type selon les valeurs d'un champ cible numérique. Comme pour les arbres décision, les arbres de régression décomposent les données en sous-ensembles dans lesquels les feuilles de l'arbre correspondent à des sous-ensembles suffisamment petits ou uniformes. Les divisions sont sélectionnées pour augmenter la dispersion des valeurs des attributs cibles, afin qu'ils soient assez prévisibles à travers leurs valeurs moyennes au niveau des feuilles.

La sortie du modèle prend la forme d'une représentation textuelle de l'arbre. Chaque ligne de texte correspond à un noeud ou à une feuille, et la marge indique le niveau de l'arbre. Pour un noeud, la condition de division est affichée; pour une feuille, l'étiquette de classe attribuée apparaît.

### **Options de création d'arbre de régression Netezza - Développement de l'arbre**

L'onglet Options de création est celui où vous définissez toutes les options de création d'un modèle. Vous pouvez bien sûr cliquer simplement sur le bouton Exécuter pour créer un modèle avec toutes les options par défaut, mais normalement vous pouvez personnaliser la création selon vos propres objectifs.

Figure 6-21  
Options de création de l'arbre de régression pour le développement de l'arbre



**Profondeur maximale de l'arbre.** Le nombre maximal de niveaux que l'arbre peut atteindre en se développant au-dessous du nœud racine (ce qui revient au nombre de fois où l'échantillon va être divisé de manière récursive). La valeur par défaut est 62, ce qui est la profondeur d'arbre maximale à des fins de modélisation. Remarquez toutefois que le visualiseur dans le nugget de modèle peut afficher un maximum de 12 niveaux.

**Critères de division.** Ces options permettent de contrôler l'arrêt de la division de l'arbre. Si vous ne voulez pas utiliser les valeurs par défaut, cliquez sur Personnaliser et effectuez les changements.

- **Mesure d'évaluation de division.** La mesure d'impureté de la classe, utilisée pour évaluer le meilleur endroit où diviser l'arbre. *Remarque* : actuellement, Variance est la seule option possible.
- **Amélioration minimale pour les divisions.** La quantité minimale d'impureté devant être réduite avant qu'une nouvelle division puisse être créée dans l'arbre. La création d'arbre vise à créer des sous-groupes ayant des résultats similaires, c'est-à-dire à minimiser l'impureté de chaque nœud. Si le meilleur découpage calculé pour une branche signifie une réduction

de l'impureté inférieure à la valeur spécifiée par le critère de division, alors le découpage ne sera pas effectué.

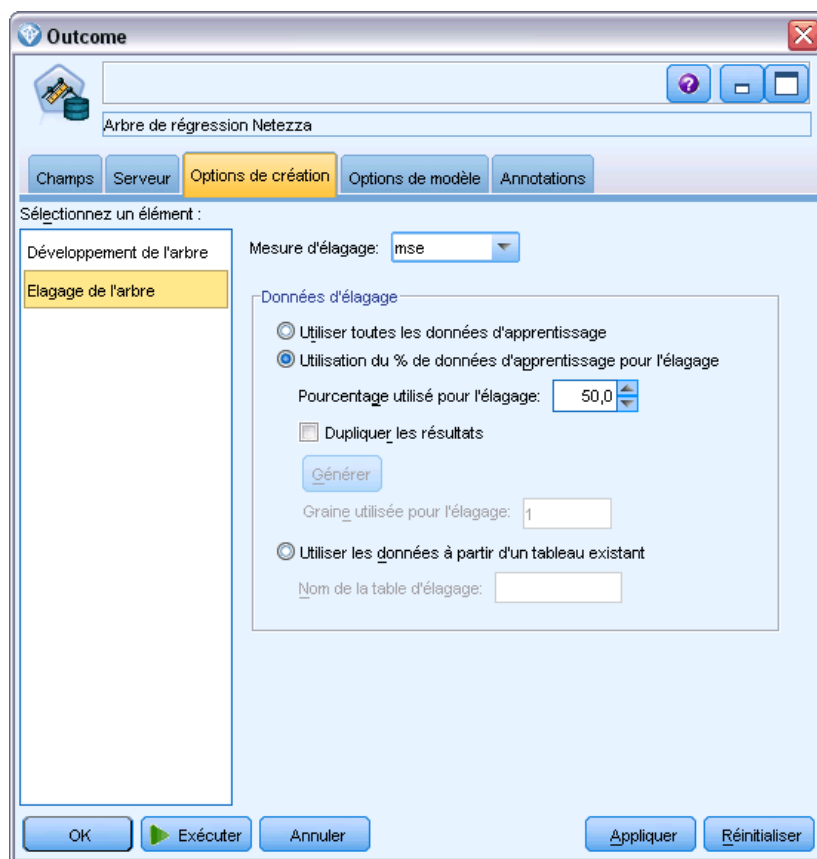
- **Nombre minimal d'instances pour une division.** Le nombre minimal d'enregistrements pouvant être divisés. Lorsqu'il reste moins d'enregistrements que ce nombre d'enregistrements non divisés, aucune division supplémentaire ne sera effectuée. Ce champ peut être utilisé pour éviter la création de sous-groupes minuscules dans l'arbre.

## Options de création d'arbre de régression Netezza - Élagage de l'arbre

Vous pouvez utiliser les options d'élagage pour spécifier les critères d'élagage de l'arbre de régression. Le but de l'élagage est de réduire le risque de surajustement en supprimant les sous-groupes trop développés qui n'améliorent pas la précision attendue des nouvelles données.

Figure 6-22

Options de création de l'arbre de régression pour l'élagage de l'arbre



**Mesure d'élagage.** La valeur de la mesure d'élagage garantit que la précision du modèle demeure dans des limites acceptables après avoir supprimé une feuille de l'arbre. Vous pouvez choisir l'une des mesures suivantes.

- **eqm.** Erreur quadratique moyenne - (par défaut) mesure à quel point une ligne intégrée est proche des points de données.

- **r2.** R2 (coefficient de détermination) - mesure la proportion de la variation de la variable dépendante, expliquée par le modèle de régression.
- **Pearson.** Coefficient de corrélation de Pearson - mesure la force de la relation entre deux variables linéairement dépendantes et normalement distribuées.
- **Spearman.** Coefficient de corrélation de Spearman - détecte les relations non linéaires paraissant faibles par rapport à la corrélation de Pearson, mais qui pourraient en fait être fortes.

**Données pour l'élagage.** Vous pouvez utiliser une partie ou la totalité des données d'apprentissage pour estimer la précision attendue des nouvelles données. Vous pouvez également utiliser un ensemble de données d'élagage distinct d'une table spécifique à cette fin.

- **Utiliser toutes les données d'apprentissage.** Cette option (option par défaut) utilise toutes les données d'apprentissage pour estimer la précision du modèle.
- **Utiliser % des données d'apprentissage pour l'élagage.** Utilisez cette option pour diviser les données en deux ensembles, un pour l'apprentissage et un pour l'élagage, à l'aide du pourcentage spécifié ici pour les données d'élagage.  
Sélectionnez Dupliquer les résultats si vous voulez spécifier une graine aléatoire pour vous assurer que les données sont partitionnées de la même façon chaque fois que vous exécutez le flux. Vous pouvez spécifier un entier dans le champ Graine utilisée pour l'élagage, ou bien cliquer sur Générer, ce qui créera un entier pseudo-aléatoire.
- **Utiliser des données d'une table existante.** Spécifiez le nom de la table d'un ensemble de données d'élagage distinct pour estimer la précision du modèle. Ceci est plus fiable que d'utiliser des données d'apprentissage. Toutefois, cette option peut provoquer la suppression d'un vaste sous-ensemble de données de l'ensemble d'apprentissage, ce qui réduit la qualité de l'arbre décision.

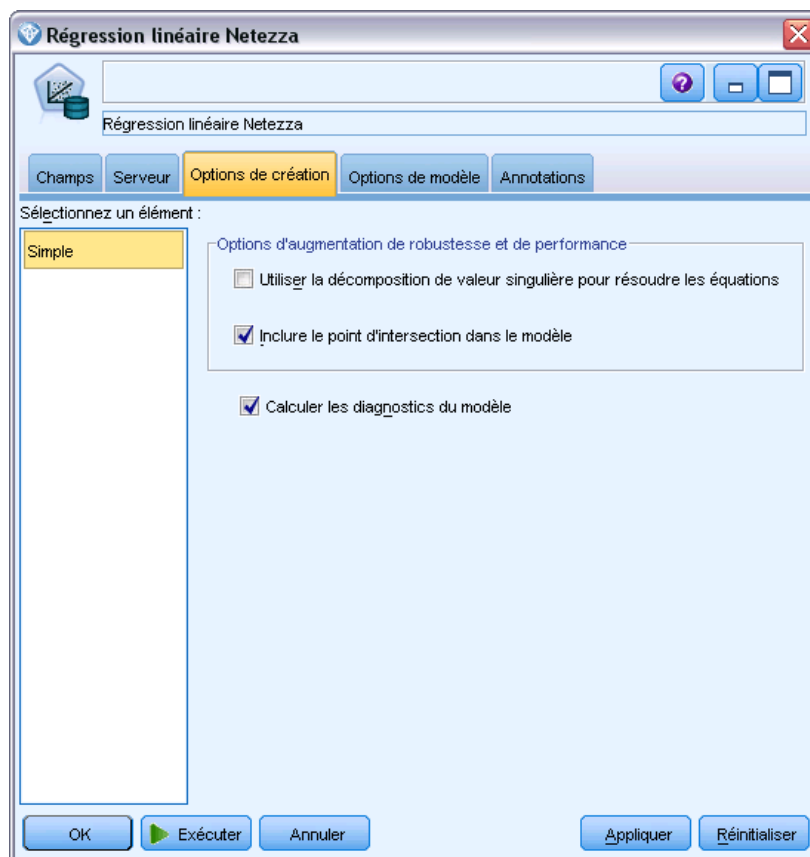
## ***Régression linéaire Netezza***

Les modèles linéaires prédisent une cible continue en fonction de relations linéaires entre la cible et une ou plusieurs variables prédites. Bien que limités aux relations linéaires de modélisation directes uniquement, les modèles de régression linéaire sont relativement simples et donnent une formule mathématique simple à interpréter pour le scoring. Les modèles linéaires sont rapides, efficaces et faciles à utiliser, bien que leur application soit limitée par rapports à ceux produits par des algorithmes de régression plus affinés.

### ***Options de création de régression linéaire Netezza***

L'onglet Options de création est celui où vous définissez toutes les options de création d'un modèle. Vous pouvez bien sûr cliquer simplement sur le bouton Exécuter pour créer un modèle avec toutes les options par défaut, mais normalement vous pouvez personnaliser la création selon vos propres objectifs.

Figure 6-23  
Options de création Régression linéaire



**Utilisez la décomposition en valeurs singulières pour résoudre les équations.** L'utilisation de la matrice de décomposition en valeurs singulières à la place de la matrice d'origine a l'avantage d'être plus résistante aux erreurs numériques et peut également accélérer le calcul.

**Inclure une constante au modèle.** L'inclusion de la constante augmente la précision globale de la solution.

**Calculer les diagnostics du modèle.** Cette option a pour effet que de nombreux diagnostics sont calculés sur le modèle. Les résultats sont stockés dans des matrices ou des tableaux pour une consultation ultérieure. Les diagnostics comprennent le  $R^2$ , la somme des carrés résiduelle, l'estimation de la variance, l'écart-type, la valeur  $p$  et la valeur  $t$ .

Ces diagnostics sont liés à la validité et à l'utilité du modèle. Vous devez exécuter des diagnostics séparés sur les données sous-jacentes afin de vous assurer qu'ils répondent aux hypothèses relatives à la linéarité.

## Séries temporelles Netezza

Une **série temporelle** est une séquence de valeurs de données numériques, mesurées à différents moments (bien que pas nécessairement régulièrement), par exemple, les prix des stocks quotidiens ou les données de vente hebdomadaires. L'analyse de ces données peut être utile, par exemple pour souligner un comportement, tel que des tendances ou la saisonnalité (un motif répétitif) et pour prévoir un comportement futur à partir d'événements passés.

Les séries temporelles Netezza prennent en charge les algorithmes de séries temporelles suivants.

- analyse spectrale
- lissage exponentiel
- AutoRegressive Integrated Moving Average (processus autorégressif moyenne mobile intégrée - ARIMA).
- décomposition des tendances saisonnières

Ces algorithmes divisent une série temporelle en un composant de tendance et un composant saisonnier. Ces composants sont ensuite analysés afin de créer un modèle pouvant être utilisé pour les prévisions.

L'**analyse spectrale** permet d'identifier un comportement périodique dans des séries temporelles. Pour les séries temporelles composées de plusieurs périodicités sous-jacentes ou lorsqu'une quantité considérable de bruit aléatoire est présente dans les données, l'analyse spectrale est le moyen le plus direct d'identifier des composants périodiques. Cette méthode détecte les fréquences de comportement périodique en transformant les séries du domaine temporel en séries du domaine fréquentiel.

Le **lissage exponentiel** est une méthode qui vise à prévoir des valeurs futures à partir de valeurs pondérées d'observations de séries antérieures. Avec le lissage exponentiel, l'influence des observations diminue avec le temps de manière exponentielle. Cette méthode prévoit un point à la fois, ajuste ses prévisions au fur et à mesure de l'arrivée de nouvelles données et prend en compte les ajouts, les tendances et la saisonnalité.

Les modèles **ARIMA** proposent des méthodes plus sophistiquées de modélisation des composants de tendance et saisonniers que les modèles de lissage exponentiel. Cette méthode suppose la spécification explicite des ordres autorégressifs et de moyennes mobiles, ainsi que du degré de différenciation.

*Remarque* : Sur le plan pratique, les modèles ARIMA s'avèrent particulièrement utiles pour inclure des valeurs prédites susceptibles d'expliquer le comportement des séries en cours de prévision, telles que le nombre de catalogues envoyés par courrier électronique ou le nombre d'accès à la page Web d'une société. Les modèles de lissage exponentiel décrivent le comportement des séries temporelles sans essayer d'en analyser les causes.

La **décomposition des tendances saisonnières** supprime les comportements périodiques des séries temporelles afin d'effectuer une analyse de tendance puis elle sélectionne une forme de base pour la tendance, telle qu'une fonction quadratique. Ces formes de base contiennent un certain nombre de paramètres dont les valeurs sont déterminées de manière à réduire l'erreur quadratique

moyenne des résidus (c'est-à-dire, les différences entre les valeurs ajustées et observées des séries temporelles).

### **Interpolation des valeurs dans les séries temporelles Netezza**

L'**interpolation** est le processus d'évaluation et d'insertion des valeurs manquantes dans les données des séries temporelles.

Si les intervalles des séries temporelles sont réguliers mais que certaines valeurs sont absentes, les valeurs manquantes peuvent être évaluées à l'aide de l'interpolation linéaire. Observons les séries suivantes d'arrivées mensuelles de passagers au terminal d'un aéroport.

**Table 6-1**

*Arrivées mensuelles au terminal des passagers*

<b>Mois</b>	<b>Passagers</b>
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

Dans ce cas, l'interpolation linéaire évalue la valeur manquante du mois 5 à hauteur de 3 650 000 (le point intermédiaire entre les mois 4 et 6).

Les intervalles irréguliers sont traités différemment. Observons les séries suivantes de relevés de température.

**Table 6-2**

*Relevés de température*

<b>Date</b>	<b>Temps</b>	<b>Température</b>
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

Les relevés sont effectués à trois moments différents pendant trois jours mais à différentes heures, dont seules certaines sont communes aux trois jours. De plus, seuls deux jours sont consécutifs.

Cette situation peut être traitée de deux façons : en calculant les agrégats ou en déterminant une taille de pas.

Les agrégats peuvent être des agrégats quotidiens calculés selon une formule basée sur la connaissance sémantique des données. Cette méthode pourrait générer l'ensemble de données suivant.

**Table 6-3**  
*Relevés de température (agrégés)*

Date	Temps	Température
2011-07-24	24:00	69
2011-07-25	24:00	71
2011-07-26	24:00	null
2011-07-27	24:00	72

L'algorithme peut également traiter la série comme une série distincte et déterminer une taille de pas appropriée. Dans ce cas, la taille de pas déterminée par l'algorithme peut être de 8 heures et peut générer le résultat suivant.

**Table 6-4**  
*Relevés de température avec taille de pas calculée*

Date	Temps	Température
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

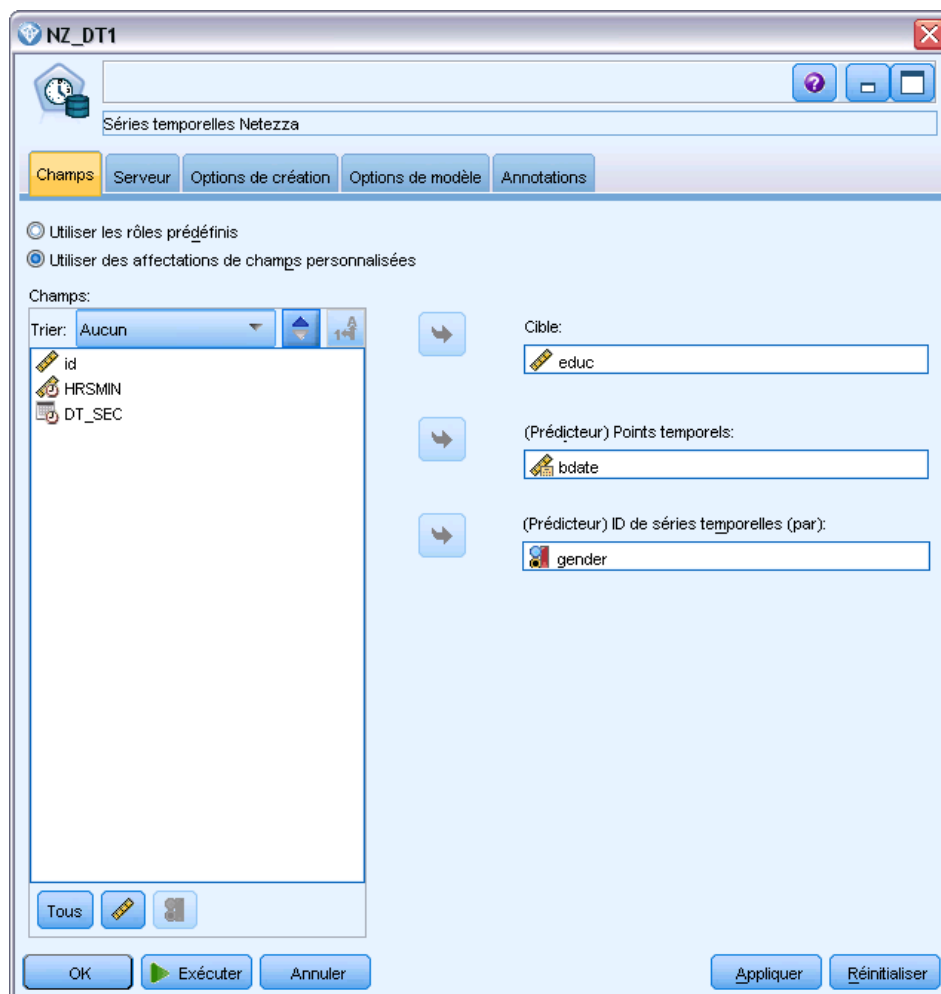
Ici, seuls quatre relevés correspondent aux mesures d'origine, mais avec l'aide d'autres valeurs connues des séries d'origine, les valeurs manquantes peuvent être de nouveau calculées par interpolation.

### ***Options des champs de séries temporelles Netezza***

Dans l'onglet Champs, spécifiez les rôles des champs d'entrée dans les données source.



Figure 6-24  
Options des champs de séries temporelles



**Champs.** Utilisez les boutons fléchés pour affecter manuellement des éléments à partir de cette liste aux divers champs de rôle à droite de l'écran. Les icônes indiquent les niveaux de mesure valides pour chaque champ de rôle. [Pour plus d'informations, reportez-vous à la section Niveaux de mesure dans le chapitre 4 dans \*Noeuds source, exécution et de sortie de IBM SPSS Modeler 15.\*](#)

**Cible.** Sélectionnez un champ comme cible pour la prédiction. Il doit s'agir d'un champ avec un niveau de mesure continu.

**(Valeurs prédites) Points temporels.** (obligatoire) Le champ d'entrée contenant les valeurs de date ou d'heure pour la série temporelle. Il doit s'agir d'un champ avec un niveau de mesure continu ou catégoriel et d'un type de stockage de données Date, Heure, Horodatage ou Numérique. Le type de stockage de données du champ que vous spécifiez ici définit également le type d'entrée de certains champs dans d'autres onglets de ce nœud de modélisation. [Pour plus d'informations, reportez-vous à la section Définition du stockage et du formatage des champs dans le chapitre 2 dans \*Noeuds source, exécution et de sortie de IBM SPSS Modeler 15.\*](#)

**(Valeur prédite) ID de séries temporelles (par).** Un champ contenant des ID de séries temporelles à utiliser si l'entrée contient plusieurs séries temporelles.

### ***Options de création de séries temporelles Netezza***

Il existe deux niveaux d'options de création :

- Basique - paramètres du choix des algorithmes, de l'interpolation et de l'intervalle de temps à utiliser.
- Avancé - paramètres des prévisions

Cette section décrit les options de base.

L'onglet Options de création est celui où vous définissez toutes les options de création d'un modèle. Vous pouvez bien sûr cliquer simplement sur le bouton Exécuter pour créer un modèle avec toutes les options par défaut, mais normalement vous pouvez personnaliser la création selon vos propres objectifs.

Figure 6-25  
Options de création de base de séries temporelles

### Algorithme

Il s'agit des paramètres associés à l'algorithme de séries temporelles utilisé.

**Nom de l'algorithme.** Choisissez l'algorithme de séries temporelles à utiliser. Les algorithmes disponibles sont des algorithmes d'analyse spectrale, de lissage exponentiel (par défaut), ARIMA ou de décomposition des tendances saisonnières. [Pour plus d'informations, reportez-vous à la section Séries temporelles Netezza sur p. 204.](#)

**Tendance.** (Lissage exponentiel uniquement) Le lissage exponentiel simple ne s'effectue pas correctement si la série temporelle présente une tendance. Utilisez ce champ pour spécifier la tendance (le cas échéant) afin que l'algorithme puisse la prendre en compte.

- **Déterminée par le système.** (par défaut) Le système essaie de trouver une valeur optimale pour ce paramètre.
- **Aucune(N).** La série temporelle n'affiche aucune tendance.

- **Additive(A)**. Une tendance qui augmente régulièrement au fur et à mesure.
- **Additive amortie(DA)**. Une tendance additive qui finit par disparaître.
- **Multiplicative(M)**. Une tendance qui augmente au fur et à mesure, généralement plus rapidement qu'une tendance additive régulière.
- **Multiplicative amortie(DM)**. Une tendance multiplicative qui finit par disparaître.

**Saisonnalité.** (Lissage exponentiel uniquement) Utilisez ce champ pour spécifier si la série temporelle présente des motifs saisonniers dans les données.

- **Déterminée par le système.** (par défaut) Le système essaie de trouver une valeur optimale pour ce paramètre.
- **Aucune(N)**. La série temporelle n'affiche aucun motif saisonnier.
- **Additive(A)**. Le motif de variations saisonnières présente une tendance croissante régulière.
- **Multiplicative(M)**. Comme la saisonnalité additive, mais dans ce cas, l'amplitude (la distance entre les points supérieurs et inférieurs) des variations saisonnières augmente en fonction de la tendance croissante générale des variations.

**Utiliser les paramètres déterminés par le système pour ARIMA.** (ARIMA uniquement) Choisissez cette option si vous souhaitez que le système détermine les paramètres pour l'algorithme ARIMA.

**Spécifier.** (ARIMA uniquement) Choisissez cette option et cliquez sur le bouton pour spécifier les paramètres ARIMA manuellement.

### **Interpolation**

Si les données source des séries temporelles contiennent des valeurs manquantes, choisissez une méthode d'insertion des valeurs estimées pour remplir les données manquantes. [Pour plus d'informations, reportez-vous à la section Interpolation des valeurs dans les séries temporelles Netezza sur p. 205.](#)

- **Linéaire.** Choisissez cette méthode si les intervalles de la série temporelle sont réguliers mais que certaines valeurs sont absentes.
- **Splines exponentielles.** Ajuste une courbe lissée à l'endroit où les valeurs des points de données connus augmentent ou diminuent à un niveau élevé.
- **Splines cubiques.** Ajuste une courbe lissée sur les points de données connus afin d'évaluer les valeurs manquantes.

### **Intervalle de temps**

C'est là que vous pouvez choisir d'utiliser toute la gamme de données dans la série temporelle ou un sous-ensemble contigu de ces données pour créer le modèle. L'entrée valide pour ces champs est définie par le type de stockage de données du champ spécifié pour les Points temporels dans l'onglet Champs. [Pour plus d'informations, reportez-vous à la section Options des champs de séries temporelles Netezza sur p. 206.](#)

- **Utiliser les premières et dernières heures disponibles dans les données.** Choisissez cette option si vous souhaitez utiliser la gamme entière des données de séries temporelles.
- **Spécifier la fenêtre temporelle.** Choisissez cette option si vous souhaitez utiliser uniquement une partie des séries temporelles. Utilisez les champs Première heure (de) et Dernière heure (à) pour spécifier les limites.

## Structure ARIMA

Figure 6-26  
Paramètres ARIMA pour les séries temporelles

Spécifiez les valeurs des différents composants saisonniers et non saisonniers du modèle ARIMA. Dans chaque cas, définissez l'opérateur sur < (inférieur à), = (égal à), ou <= (inférieur ou égal à), puis spécifiez la valeur dans le champ adjacent. Les valeurs doivent être des entiers positifs spécifiant les degrés.

**Non saisonnières.** Les valeurs des différents composants non saisonniers du modèle.

- **Degrés d'autocorrélation (p).** Le nombre d'ordres autorégressifs dans le modèle. Les ordres autorégressifs indiquent quelles valeurs précédentes de la série seront utilisées pour prévoir les valeurs en cours. Par exemple, un ordre autorégressif de 2 indique que la valeur de la série Deux points dans le temps dans le passé sera utilisée pour prévoir la valeur en cours.
- **Calcul (d).** Spécifie l'ordre de différenciation appliqué à la série avant d'estimer les modèles. La différenciation est nécessaire lorsque les tendances sont présentes (les séries avec tendances sont en général non stationnaires et la modélisation ARIMA suppose la stationnarité) et est utilisée pour supprimer leurs effets. L'ordre de différenciation correspond au degré de tendance de série, aux comptes de différenciation de premier ordre pour les tendances linéaires, aux comptes de différenciation de second ordre pour les tendances quadratiques, etc.
- **Moyenne mobile (q).** Le nombre d'ordres de moyenne mobile dans le modèle. Les ordres de moyenne mobile indiquent comment les écarts de la moyenne de la série pour les valeurs précédentes sont utilisés pour prévoir les valeurs courantes. Par exemple, des ordres de moyenne mobile de 1 et 2 indiquent que les écarts de la valeur de la moyenne de la série pour chacune des deux dernières périodes doivent être considérés lors de la prévision des valeurs actuelles de la série.

**Saisonnier.** Les composants d'autocorrélation saisonnière (SP), de calcul saisonnier (DS) et de moyenne mobile saisonnière (SQ) jouent les mêmes rôles que leurs équivalents non saisonniers. Cependant, pour les ordres saisonniers, les valeurs courantes de la série sont affectées par les valeurs de série précédentes séparées par une ou plusieurs périodes saisonnières. Par exemple, pour des données mensuelles (période saisonnière de 12), un ordre saisonnier de 1 indique que la valeur de série en cours est affectée par les 12 périodes de la valeur de série précédant celle en cours. Un ordre saisonnier de 1, pour des données mensuelles, est alors le même que lorsqu'on spécifie un ordre non saisonnier de 12.

Les paramètres saisonniers sont pris en compte uniquement si la saisonnalité est détectée dans les données ou si vous spécifiez les paramètres Période dans l'onglet Avancé.

### Options de création de séries temporelles Netezza - Niveau avancé

Vous pouvez utiliser les paramètres avancés pour spécifier les options des prévisions.

Figure 6-27  
Options de création avancées de séries temporelles

The screenshot shows a dialog box titled 'NZ\_DT1' with a tabbed interface. The 'Options de création' tab is selected. On the left, there is a sidebar with 'Basic' and 'Advanced' tabs, where 'Advanced' is highlighted. The main area is titled 'Séries temporelles Netezza' and contains the following options:

- Utilisez les paramètres déterminés par le système pour les options de construction de modèles
- Spécifier
  - Période: 5 (spin box)
  - Unités de période: Jours (dropdown menu)
  - Horizon prévisionnel
    - 1999-12-12 (text box)
    - YYYY-MM-DD (format label)
  - Heures prévisionnelles
    - Entrées d'heures prévisionnelles (text box)
    - 1921-03-02 (text box)
    - YYYY-MM-DD (format label)

At the bottom, there are buttons for 'OK', 'Exécuter', 'Annuler', 'Appliquer', and 'Réinitialiser'.

**Utiliser les paramètres déterminés par le système pour les options de création de modèles.** Choisissez cette option si vous souhaitez que le système détermine les paramètres avancés.

**Spécifier.** Choisissez cette option si vous souhaitez spécifier les options avancées manuellement. (Cette option n'est pas disponible si l'algorithme est une analyse spectrale).

- **Période/Unités pour la période.** La période de temps après laquelle un comportement caractéristique de séries temporelles se répète. Par exemple, pour une série temporelle de résultats de vente hebdomadaires, vous indiqueriez 1 pour la période et Semaines pour les unités. La Période doit être un entier positif ; les Unités pour la période peuvent être des millisecondes, secondes, minutes, heures, jours, semaines, trimestres ou années. Ne définissez pas les Unités pour la période si Période n'est pas défini ou si le type de temps n'est pas numérique. Cependant, si vous spécifiez la Période, vous devez également spécifier les Unités pour la période.

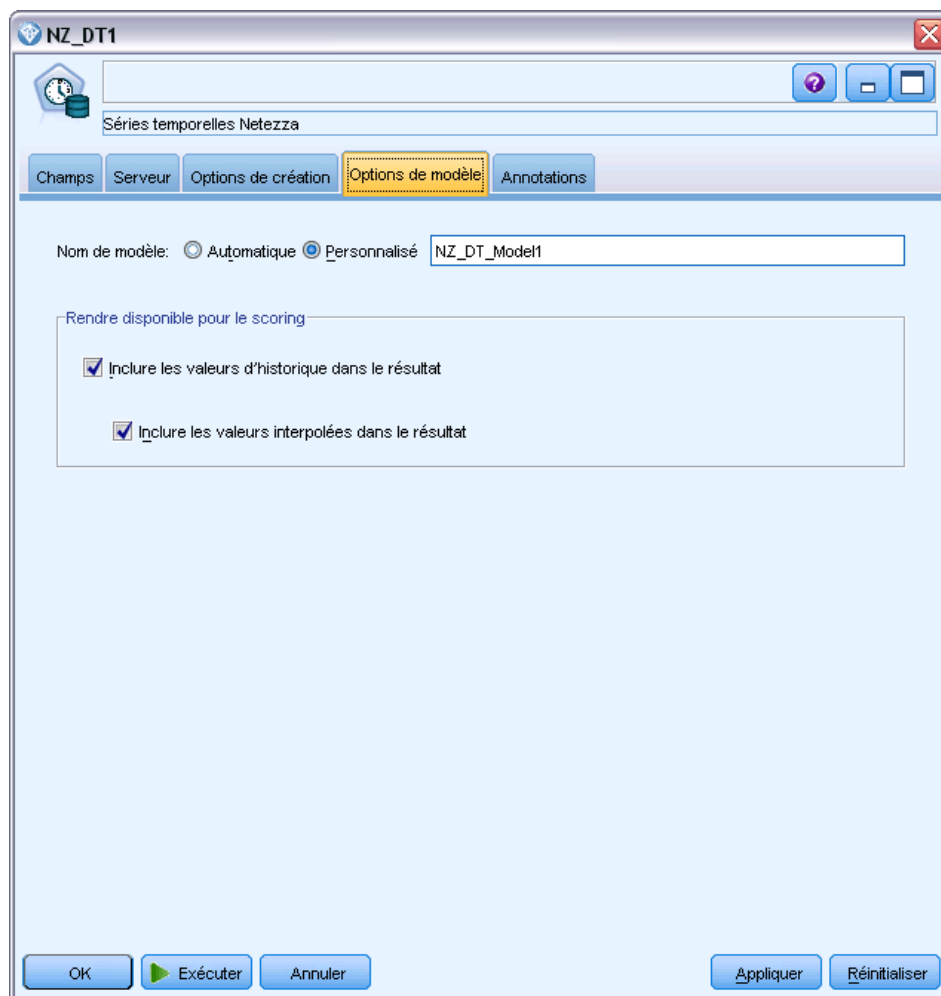
**Paramètres des prévisions.** Vous pouvez choisir d'effectuer des prévisions jusqu'à un certain point temporel ou à des points temporels spécifiques. L'entrée valide pour ces champs est définie par le type de stockage de données du champ spécifié pour les Points temporels dans l'onglet Champs. [Pour plus d'informations, reportez-vous à la section Options des champs de séries temporelles Netezza sur p. 206.](#)

- **Horizon prévisionnel.** Choisissez cette option si vous souhaitez uniquement spécifier un point final pour les prévisions. Les prévisions seront effectuées jusqu'à ce point temporel.
- **Heures des prévisions.** Choisissez cette option pour spécifier un ou plusieurs points temporels auxquels effectuer des prévisions. Cliquez sur Ajouter pour ajouter une nouvelle ligne à la table des points temporels. Pour supprimer une ligne, sélectionnez-la et cliquez sur Supprimer.

## ***Options du modèle de série temporelle Netezza***

Dans l'onglet Options de modèle, vous pouvez choisir de spécifier un nom pour le modèle ou de générer automatiquement un nom. Vous pouvez également définir les valeurs par défaut des options de sortie du modèle.

Figure 6-28  
Options de modèle de séries temporelles



**Nom du modèle.** Vous pouvez générer le nom du modèle automatiquement sur la base du champ cible ou ID (ou du type de modèle si aucun de ces champs n'est spécifié) ou spécifier un nom personnalisé.

**Rendre disponible pour le scoring.** Vous pouvez définir ici les valeurs par défaut pour les options de scoring qui apparaissent dans la boîte de dialogue du nugget de modèle.

- **Inclure les valeurs historiques dans le résultat.** Par défaut, la sortie du modèle ne contient pas les valeurs des données historiques (celles utilisées pour effectuer la prévision). Sélectionnez cette case pour inclure ces valeurs.
- **Inclure les valeurs interpolées dans le résultat.** Si vous choisissez d'inclure les valeurs historiques dans la sortie, sélectionnez cette case si vous souhaitez également inclure les valeurs interpolées, le cas échéant. Veuillez noter que l'interpolation fonctionne uniquement sur des données historiques. Par conséquent, cette case n'est pas disponible si Inclure les valeurs historiques dans le résultat n'est pas sélectionné. [Pour plus d'informations,](#)



reportez-vous à la section [Interpolation des valeurs dans les séries temporelles Netezza](#) sur p. 205.

## **Linéaire généralisé Netezza**

La régression linéaire est une technique statistique ancienne de classification des enregistrements sur la base des valeurs des champs d'entrée numériques. La régression linéaire correspond à une ligne droite ou à une surface qui minimise les écarts entre les valeurs prédites et les valeurs réelles des résultats. Les modèles linéaires sont utiles pour la modélisation d'une large gamme de phénomènes réels en raison de leur simplicité à la fois lors de l'application de l'apprentissage et des modèles. Cependant, les modèles linéaires supposent une distribution normale dans la variable (cible) dépendante et un impact linéaire des variables indépendantes (valeurs prédites) sur la variable dépendante.

Il existe de nombreux cas où une régression linéaire est utile mais où les suppositions précédentes ne s'appliquent pas. Par exemple, lors de la modélisation du choix des consommateurs par rapport à un nombre peu élevé de produits, la variable dépendante est susceptible d'avoir une distribution multinomiale. De même, lors de la modélisation des revenus par rapport à l'âge, ces revenus augmentent généralement parallèlement à l'âge, mais le lien entre ces deux éléments est peu susceptible d'être aussi simple qu'une ligne droite.

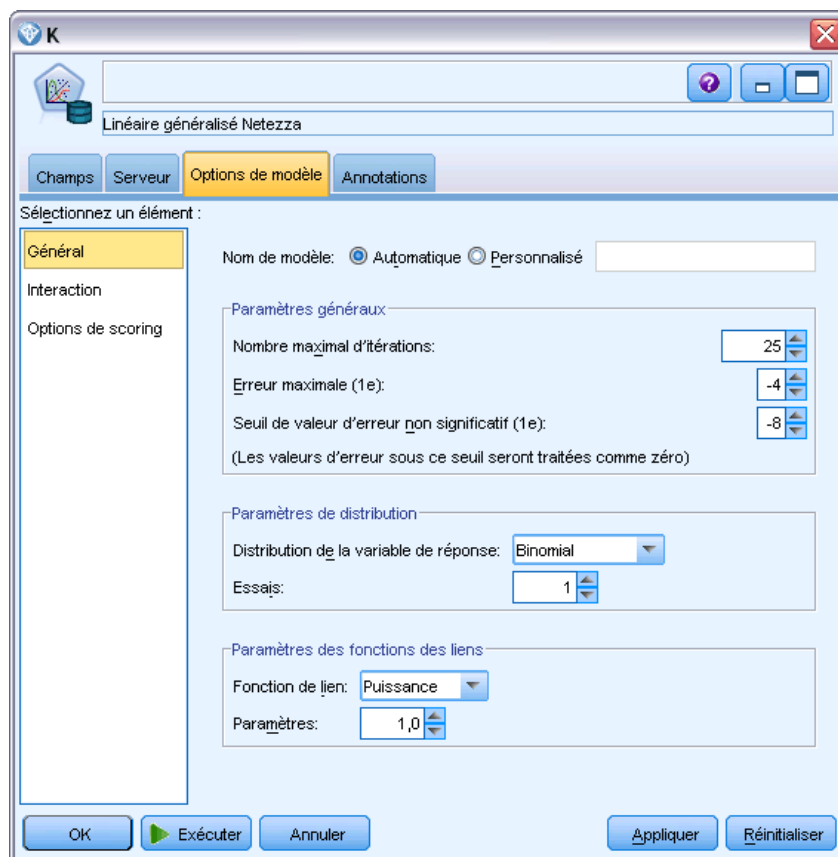
Dans ce genre de cas, un modèle linéaire généralisé peut être utilisé. Les modèles linéaires généralisés développent le modèle de régression linéaire afin que la variable dépendante soit en lien avec les variables prédictives au moyen d'une fonction de liaison spécifique pour laquelle plusieurs fonctions appropriées existent. De plus, le modèle permet à la variable dépendante d'avoir une distribution anormale, telle qu'une distribution de Poisson, binomiale, etc.

L'algorithme recherche le modèle le plus adapté de manière répétée, jusqu'à un nombre d'itérations donné. Lors du calcul du modèle le plus adapté, l'erreur est représentée par la somme des carrés des différences entre la valeur prédite et la valeur réelle de la variable dépendante.

### **Options de modèle linéaire généralisé Netezza - Généralités**

Dans l'onglet Options de modèle, vous pouvez choisir de spécifier un nom pour le modèle ou de générer automatiquement un nom. Vous pouvez également choisir plusieurs paramètres associés au modèle, à la fonction de lien, aux interactions des champs d'entrée (le cas échéant) et définir les valeurs par défaut des options de scoring.

Figure 6-29  
Options générales de modèle linéaire généralisé



**Nom du modèle.** Vous pouvez générer le nom du modèle automatiquement sur la base du champ cible ou ID (ou du type de modèle si aucun de ces champs n'est spécifié) ou spécifier un nom personnalisé.

**Paramètres généraux.** Ces paramètres sont associés aux critères d'arrêt de l'algorithme.

- **le nombre maximum des itérations.** Le nombre maximum d'itérations que l'algorithme exécutera ; le minimum est de 1, la valeur par défaut est de 20.
- **Nombre maximale d'erreurs (1e).** La valeur du nombre maximal d'erreurs (en notation scientifique) à laquelle l'algorithme doit s'arrêter de rechercher le modèle le plus adapté. Le minimum est de 0, la valeur par défaut est de -3, ce qui signifie 1E-3 ou 0,001.
- **Seuil de valeurs d'erreurs non significatives (1e).** La valeur (en notation scientifique) sous laquelle les erreurs sont traitées comme ayant une valeur de zéro. Le minimum est de -1, la valeur par défaut est de -7 ce qui signifie que les valeurs d'erreur inférieures à 1E-7 (ou 0,0000001) seront comptabilisées comme non significatives.

**Paramètres de distribution.** Ces paramètres sont associés à la distribution de la variable (cible) dépendante

- **Distribution de la variable de réponse.** Le type de distribution est au choix de Bernoulli (par défaut), Gaussienne, de Poisson, Binomiale, Binomiale négative, Wald (gaussienne inversée) et Gamma.
- **Essais.** (Distribution binomiale uniquement, auquel cas elle est requise). Lorsque la réponse cible est un nombre d'événements se produisant dans un ensemble d'essais, le champ cible contient le nombre d'événements et le champ Essais contient le nombre d'essais. Par exemple, lors du test d'un nouveau pesticide, vous devez exposer des échantillons de fourmis à différentes concentrations de pesticide et enregistrer le nombre de fourmis tuées et le nombre de fourmis exposées dans chaque échantillon. Dans ce cas, le champ enregistrant le nombre de fourmis tuées doit être spécifié comme le champ cible (d'événements), et le champ enregistrant le nombre de fourmis présentes dans chaque échantillon doit être spécifié comme le champ d'essais. Le nombre d'essais doit être un entier positif supérieur ou égal au nombre d'événements pour chaque enregistrement.
- **Paramètres.** (Distribution binomiale négative uniquement) Vous pouvez spécifier une valeur de paramètre si la distribution est binomiale négative. Choisissez une valeur ou utilisez la valeur par défaut de -1.

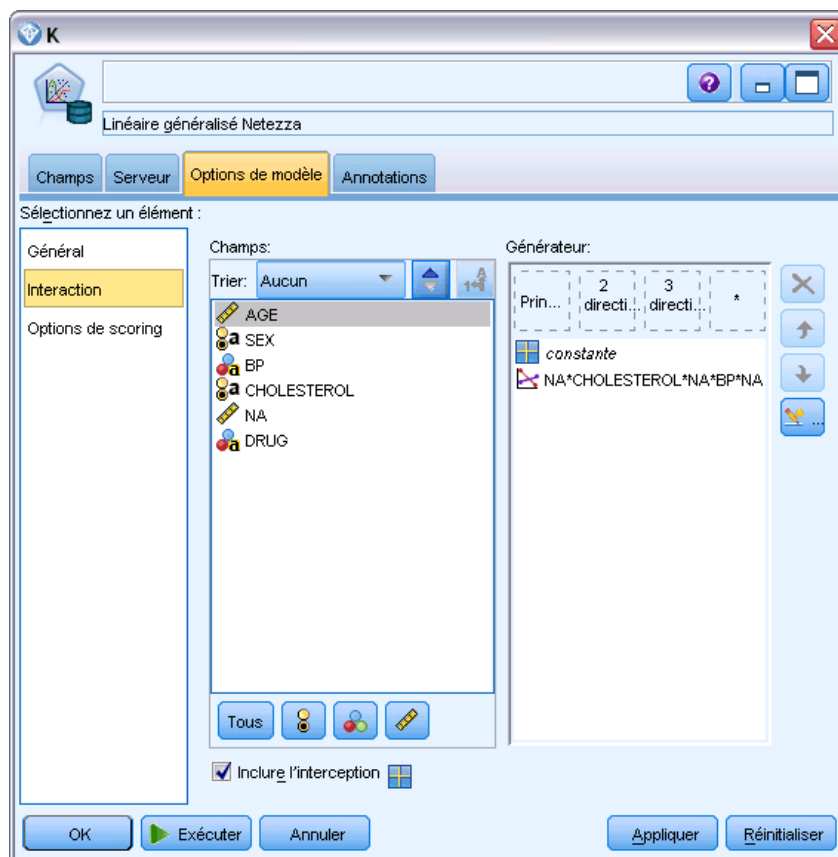
**Paramètres des fonctions de lien.** Ces paramètres sont associés à la fonction de lien qui associe la variable dépendante aux variables prédictives.

- **Fonction de lien.** La fonction à utiliser est au choix Identity, Inverse, Invnegative, Invsquare, Sqrt, Power, Oddspower, Log, Clog, Loglog, Cloglog, Logit (par défaut), Probit, Gaussit, Cauchit, Canbinom, Cangeom, Cannegbinom.
- **Paramètres.** (Fonctions de lien Power ou Oddspower uniquement). Vous pouvez spécifier une valeur de paramètre si la fonction de lien est Power ou Oddspower. Choisissez une valeur ou utilisez la valeur par défaut de 1.

### ***Options de modèle linéaire généralisé Netezza - Interactions***

Le volet Interactions contient les options de spécification des interactions (c'est-à-dire les effets multiplicatifs entre les champs d'entrée).

Figure 6-30  
Options d'interactions de modèle linéaire généralisé



**Interaction des colonnes.** Sélectionnez cette case pour spécifier les interactions entre les champs d'entrée. Ne cochez pas cette case s'il n'y a aucune interaction.

Entrez les interactions dans le modèle en sélectionnant un ou plusieurs champs dans la liste source et en les faisant glisser vers la liste des interactions. Le type d'interaction créé dépend de l'endroit où vous déposez la sélection.

- **Principaux.** Les champs déposés apparaissent sous forme d'interactions principales distinctes au bas de la liste des interactions.
- **Bidirectionnels.** Toutes les paires possibles des champs déposés apparaissent sous forme d'interactions bidirectionnelles au bas de la liste des interactions.
- **Tridirectionnels.** Tous les triplets possibles des champs déposés apparaissent sous forme d'interactions tridirectionnelles au bas de la liste des interactions.
- **\***. La combinaison de tous les champs déposés apparaît sous forme d'une unique interaction au bas de la liste des interactions.

Les boutons situés à droite de l'écran vous permettent de :



Supprimer des termes du modèle en sélectionnant ceux que vous souhaitez supprimer, puis en cliquant sur le bouton de suppression.

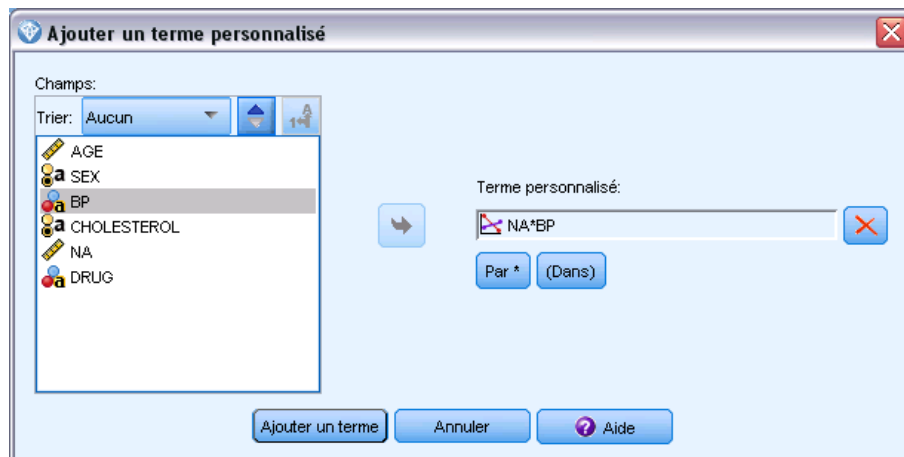


Réorganiser les termes dans le modèle, en sélectionnant ceux que vous souhaitez réorganiser, puis en cliquant sur les flèches vers le haut ou vers le bas

**Inclure la constante.** La constante est généralement incluse dans le modèle. Si vous partez du principe que les données passent par l'origine, vous pouvez exclure la constante.

### Ajouter un terme personnalisé

Figure 6-31  
Boîte de dialogue Ajouter un Terme personnalisé



Vous pouvez spécifier les interactions personnalisées sous la forme  $n1 * x1 * x1 * x1 \dots$ . Sélectionnez un champ dans la liste Champs, cliquez sur la flèche de droite pour ajouter le champ à Terme personnalisé, cliquez sur Par\*, sélectionnez le champ suivant, cliquez sur la flèche de droite, etc. Lorsque vous avez créé l'interaction personnalisée, cliquez sur Ajouter un terme pour la renvoyer dans le volet Interactions.

## **Options de modèle linéaire généralisé Netezza - Options de scoring**

**Rendre disponible pour le scoring.** Vous pouvez définir ici les valeurs par défaut pour les options de scoring qui apparaissent dans la boîte de dialogue du nugget de modèle. [Pour plus d'informations, reportez-vous à la section Nugget de modèle linéaire généralisé Netezza - Onglet Paramètres sur p. 237.](#)

- **Inclure les champs d'entrée.** Sélectionnez cette case si vous souhaitez afficher les champs d'entrée dans la sortie du modèle ainsi que les prévisions.

## **Gestion des modèles IBM Netezza Analytics**

Les modèles IBM® Netezza® Analytics sont ajoutés à l'espace de travail et à la palette Modèles comme les autres modèles IBM® SPSS® Modeler et peuvent être utilisés de la même façon. Toutefois, on distingue quelques différences importantes, dans la mesure où chaque modèle Netezza Analytics créé dans SPSS Modeler fait en réalité référence à un modèle stocké sur un serveur de base de données. Par conséquent, pour qu'un flux fonctionne correctement, il doit être connecté à une base de données dans laquelle le modèle a été créé et la table de modèles ne doit pas avoir été modifiée par un processus externe.

## **Détermination de scores des modèles IBM Netezza Analytics**

Les modèles sont représentés dans l'espace de travail par une icône de nugget de modèle dorée. L'objectif principal d'un nugget est le scoring de données afin de générer des prédictions ou de permettre une analyse ultérieure des propriétés du modèle. Les scores sont ajoutés sous la forme d'un ou plusieurs champs de données supplémentaires pouvant être rendus visibles en joignant un nœud Table au nugget et en exécutant cette branche du flux, comme décrit plus loin dans cette section. Certaines boîtes de dialogue de nugget, telles que celles de l'arbre de décision ou l'arbre de régression, ont également un onglet Modèle fournissant une représentation visuelle du modèle.

Les champs supplémentaires sont marqués par le préfixe \$<id>- ajouté au nom du champ cible, où <id> dépend du modèle et identifie le type d'information ajoutée. Les différents identificateurs sont décrits dans les rubriques de chaque nugget de modèle.

Pour afficher les scores, effectuez les étapes suivantes :

- ▶ Liez un nœud Table au nugget de modèle.
- ▶ Ouvrez le nœud Table.
- ▶ Cliquez sur Exécuter.
- ▶ Faites défiler vers la droite la fenêtre de sortie du tableau pour afficher les champs supplémentaires et leurs scores.

## Onglet serveur du nugget de modèle Netezza

Dans l'onglet Serveur, vous pouvez définir les options de serveur pour le scoring du modèle. Vous pouvez continuer à utiliser une connexion au serveur spécifiée en amont, ou vous pouvez transférer les données vers une autre base de données que vous spécifiez ici.

Figure 6-32  
Exemples d'options de serveur du nugget de modèle Netezza



**Informations sur le serveur Netezza DB.** C'est ici que vous spécifiez les informations de connexion pour la base de données à utiliser pour le modèle.

- **Utiliser la connexion en amont.** (par défaut) Utilise les informations de connexion spécifiées dans un noeud en amont, par exemple le noeud source Base de données. *Remarque* : cette option fonctionne uniquement si tous les noeuds en amont peuvent utiliser les répercussions SQL. Dans ce cas, il est inutile de déplacer les données en-dehors de la base de données car le SQL implémente entièrement tous les noeuds en amont.
- **Déplacer les données vers la connexion.** Déplace les données vers la base de données spécifiée ici. Ceci permet de faire fonctionner la modélisation si les données se trouvent dans une autre base de données IBM Netezza ou une base de données d'un autre fabricant, ou même si les données se trouvent dans un fichier plat. De plus, les données sont de nouveau déplacées vers la base de données spécifiée ici si les données ont été extraites parce qu'un noeud n'a pas effectué de répercussions SQL. Cliquez sur le bouton Edit pour rechercher et sélectionner une connexion. *Avertissement* : IBM® Netezza® Analytics est généralement utilisé avec de très grands ensembles de données. Le transfert de grandes quantités de données entre des bases de

données, ou hors d'une base de données puis dans cette même base de données, peut prendre beaucoup de temps et doit être évité aussi souvent que possible.

**Nom de la table.** Le nom de la table de base de données où le modèle est stocké. Ceci est à titre d'information uniquement ; vous ne pouvez pas modifier ce nom ici.

### ***Nuggets de modèle Arbre de décision Netezza***

Le nugget du modèle d'arbre de décision affiche le résultat de l'opération de modélisation et permet également de définir des options de scoring de modèle.

Quand vous exécutez un flux contenant un nœud de modélisation Arbre décision, le nœud ajoute un nouveau champ par défaut, dont le nom est dérivé du nom du modèle.

Table 6-5

*Champ de scoring des modèles pour l'arbre décision*

Nom du champ ajouté	Signification
<i>\$I-nom_du_modèle</i>	Valeur prédite pour l'enregistrement actuel.

Si vous avez sélectionné l'option Calcule les possibilités de classes affectées pour des enregistrements de scoring sur le nœud de modélisation ou le nugget du modèle et exécuté le flux, un champ supplémentaire est ajouté.

Table 6-6

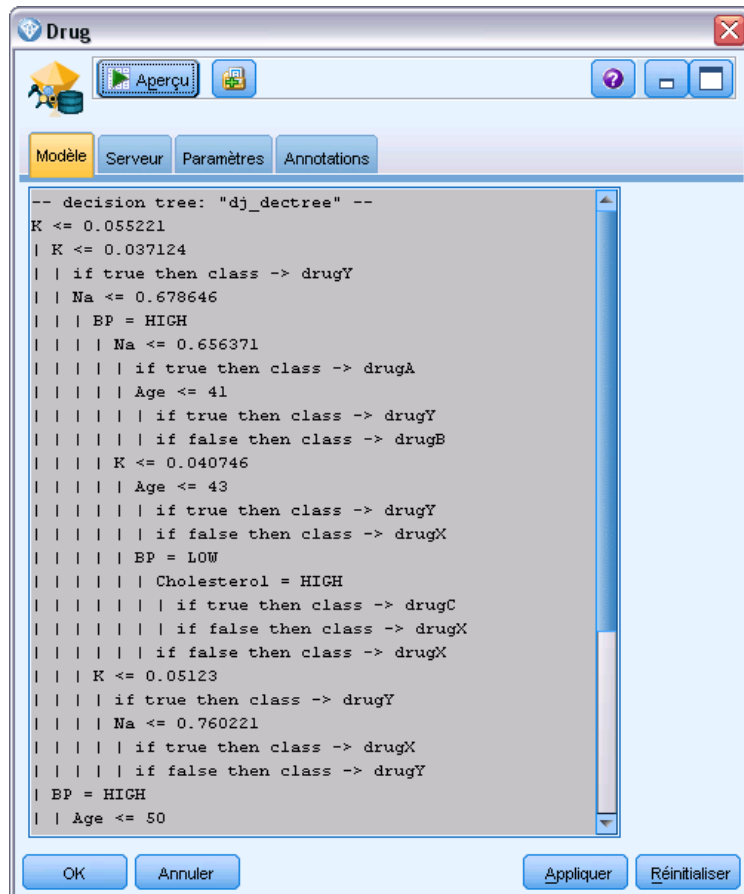
*Champ de scoring des modèles pour l'arbre décision - supplémentaire*

Nom du champ ajouté	Signification
<i>\$IP-nom_du_modèle</i>	Valeur de confiance (de 0,0 à 1,0) de la prévision.



### Nugget Arbre décision Netezza - Onglet Modèle

Figure 6-33  
Sortie du modèle Arbre décision



La sortie du modèle prend la forme d'une représentation textuelle de l'arbre. Chaque ligne de texte correspond à un noeud ou à une feuille, et la marge indique le niveau de l'arbre. Pour un noeud, la condition de division est affichée; pour une feuille, l'étiquette de classe attribuée apparaît.

### Nugget Arbre décision Netezza - Onglet Paramètres

L'onglet Paramètres vous permet de définir des options de scoring pour le modèle.

**Inclure les champs d'entrée.** Si cette option est sélectionnée, elle transmet tous les champs d'entrée originaux en aval, ajoutant le ou les champs de modélisation supplémentaires à chaque ligne de données. Si vous désélectionnez cette case, seul le champ ID d'enregistrement et les champs de modélisation supplémentaires sont transmis et par conséquent, le flux s'exécute plus rapidement.

**Calcule les probabilités de classes affectées pour des enregistrements de scoring.** (Arbre de décision et Naive Bayes uniquement) Si cette option est sélectionnée, cela signifie que les champs de modélisation supplémentaires comprennent un champ de confiance (c'est-à-dire un champ de

probabilité) ainsi qu'un champ de prévision. Si vous désélectionnez cette case, seul le champ de prévision est généré.

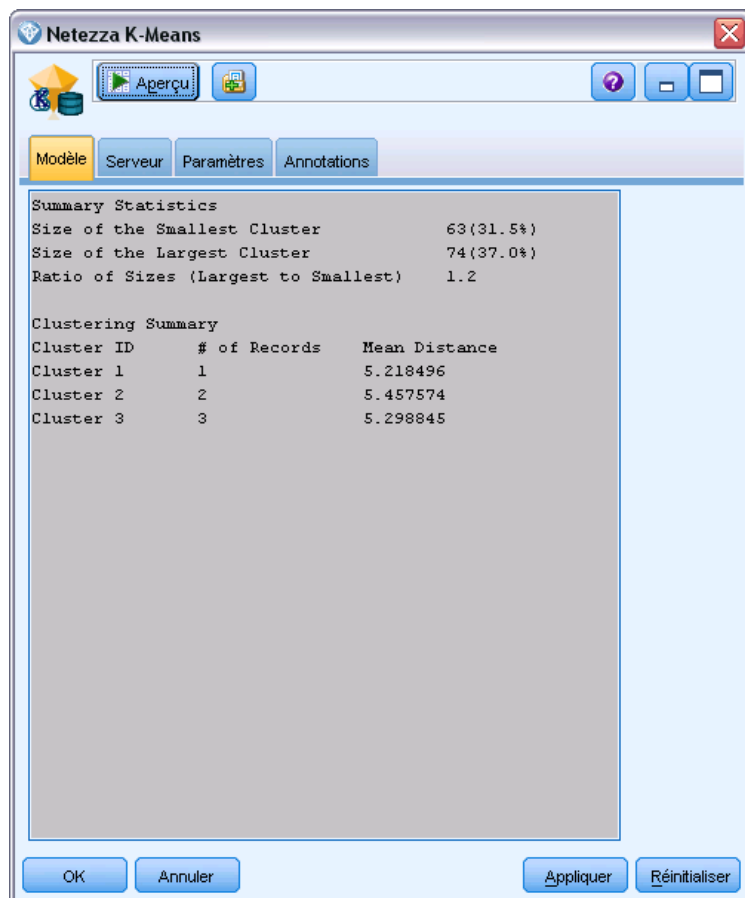
### **Nugget du modèle K-Means Netezza**

Les nuggets de modèles K-Means contiennent toutes les informations rassemblées par le modèle de classification, ainsi que des informations sur les données d'apprentissage et le processus d'estimation.

Lorsque vous exécutez un flux contenant un noeud de modèle K-Means, le noeud crée deux champs contenant la classe d'appartenance de l'enregistrement et sa distance par rapport au centre de la classe à laquelle il a été affecté. Le nom du champ contenant la classe d'appartenance est constitué du nom du modèle auquel le préfixe *\$KM-* est ajouté, et le nom du champ contenant la distance par rapport au centre de la classe est constitué du nom du modèle auquel le préfixe *\$KMD-* est ajouté. Par exemple, si le nom de votre modèle est *Kmeans*, les nouveaux champs s'intituleront *\$KM-Kmeans* et *\$KMD-Kmeans*.

### **Nugget K-Means Netezza - Onglet Modèle**

Figure 6-34  
Sortie du modèle K-means



La sortie du modèle apparaît dans l'onglet Modèle comme suit.

**Statistiques récapitulatives.** Indique le nombre d'enregistrements et le pourcentage de l'ensemble de données de la classe la plus petite et de la classe la plus grande. La liste répertorie également le rapport de taille de la plus grande et de la plus petite classe.

**Récapitulatif des classes.** Répertorie les classes créées par l'algorithme. Pour chaque classe, la table indique le nombre d'enregistrements de cette classe, ainsi que la distance moyenne depuis le centre de la classe pour ces enregistrements.

### ***Nugget K-Means Netezza - Onglet Paramètres***

L'onglet Paramètres vous permet de définir des options de scoring pour le modèle.

**Inclure les champs d'entrée.** Si cette option est sélectionnée, elle transmet tous les champs d'entrée originaux en aval, ajoutant le ou les champs de modélisation supplémentaires à chaque ligne de données. Si vous désélectionnez cette case, seul le champ ID d'enregistrement et les champs de modélisation supplémentaires sont transmis et par conséquent, le flux s'exécute plus rapidement.

**Mesure de distance.** La méthode à utiliser pour la mesure de distance entre les points de données ; plus les distances sont importantes plus les différences le sont aussi. Les options sont les suivantes :

- **Euclidienne.** (par défaut) La distance entre deux points est calculée en les reliant par une ligne droite.
- **Manhattan.** La distance entre deux points est calculée en tant que total des différences absolues entre leur coordonnées.
- **Canberra.** Semblable à la distance Manhattan, mais plus sensible aux points de données plus proches de l'origine.
- **Maximum :** la distance entre deux points est calculée comme la plus importante de leurs différences parmi toutes les dimensions de leurs coordonnées.

### ***Nuggets de modèle Bayes Net Netezza***

Le nugget de modèle Bayes Net permet de définir les options de scoring du modèle.

Quand vous exécutez un flux contenant un nœud de modélisation Bayes Net, le nœud ajoute un nouveau champ, dont le nom est dérivé du nom du modèle.

Table 6-7

*Champ de scoring des modèles pour Bayes Net*

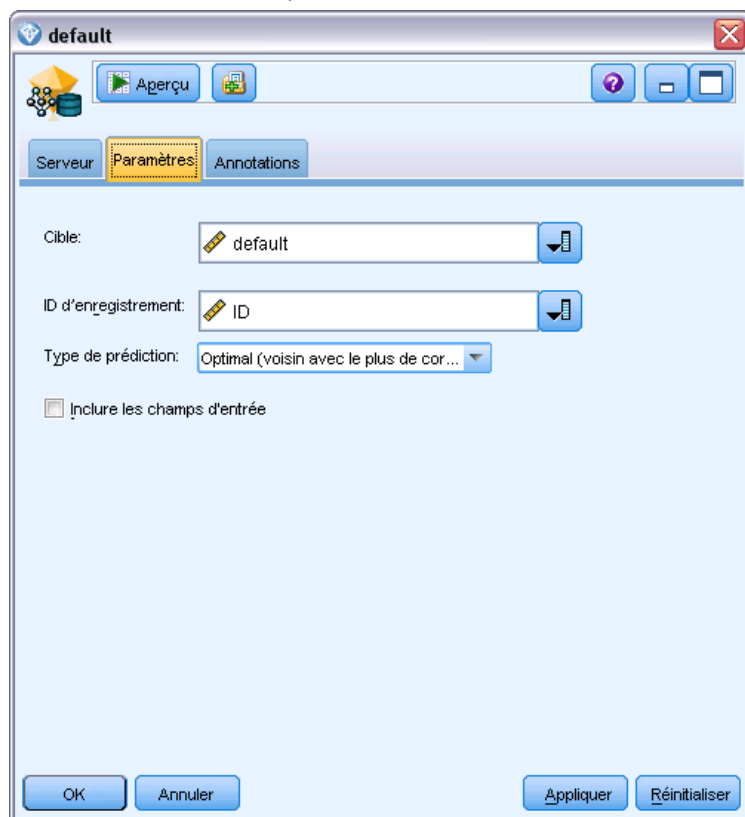
Nom du champ ajouté	Signification
\$BN-nom_du_modèle	Valeur prédite pour l'enregistrement actuel.

Vous pouvez visualiser le champ supplémentaire en joignant un nœud Table au nugget de modèle et en exécutant le nœud Table. [Pour plus d'informations, reportez-vous à la section Détermination de scores des modèles IBM Netezza Analytics sur p. 220.](#)

### **Nugget Bayes Net Netezza - Onglet Paramètres**

Dans l'onglet Paramètres, vous pouvez définir les options pour le scoring du modèle.

Figure 6-35  
Paramètres de modèle Bayes Net



**Cible.** Si vous souhaitez évaluer un champ cible différent de la cible actuelle, choisissez la nouvelle cible ici.

**ID d'enregistrement.** Si aucun champ ID d'enregistrement n'est spécifié, choisissez le champ à utiliser ici.

**Type de prévision.** La variation de l'algorithme de prévision que vous souhaitez utiliser :

- **Meilleur (voisin le plus corrélé).** (par défaut) Utilise le nœud voisin le plus corrélé.
- **Voisins (prévision pondérée des voisins).** Utilise une prévision pondérée de tous les nœuds voisins.
- **Voisins NN (voisins non nuls).** Semblable à l'option précédente, excepté qu'elle ignore les nœuds à valeur nulle (les nœuds correspondant aux attributs ayant des valeurs manquantes pour l'instance pour laquelle la prévision est calculée).

**Inclure les champs d'entrée.** Si cette option est sélectionnée, elle transmet tous les champs d'entrée originaux en aval, ajoutant le ou les champs de modélisation supplémentaires à chaque ligne de données. Si vous désélectionnez cette case, seul le champ ID d'enregistrement et les champs de modélisation supplémentaires sont transmis et par conséquent, le flux s'exécute plus rapidement.

## Nuggets de modèle Naive Bayes Netezza

Le nugget de modèle Naive Bayes permet de définir les options de scoring du modèle.

Quand vous exécutez un flux contenant un nœud de modélisation Naive Bayes, le nœud ajoute un nouveau champ par défaut, dont le nom est dérivé du nom du modèle.

Table 6-8  
Champ de scoring des modèles pour Naive Bayes - défaut

Nom du champ ajouté	Signification
\$I-nom_du_modèle	Valeur prédite pour l'enregistrement actuel.

Si vous avez sélectionné l'option Calcule les possibilités de classes affectées pour des enregistrements de scoring sur le nœud de modélisation ou le nugget du modèle et exécuté le flux, deux champs supplémentaires sont ajoutés.

Table 6-9  
Champs de scoring des modèles pour Naive Bayes - supplémentaire

Nom du champ ajouté	Signification
\$IP-nom_du_modèle	Le numérateur bayésien de la classe de l'instance (le produit de la possibilité de classe précédente et les possibilités de valeur d'attribut de l'instance conditionnelle).
\$ILP-nom_du_modèle	L'algorithme naturel de ce dernier.

Vous pouvez visualiser les champs supplémentaires en joignant un nœud Table au nugget de modèle et en exécutant le nœud Table. [Pour plus d'informations, reportez-vous à la section Détermination de scores des modèles IBM Netezza Analytics sur p. 220.](#)

## Nugget Naive Bayes Netezza - Onglet Paramètres

Dans l'onglet Paramètres, vous pouvez définir les options pour le scoring du modèle.

**Inclure les champs d'entrée.** Si cette option est sélectionnée, elle transmet tous les champs d'entrée originaux en aval, ajoutant le ou les champs de modélisation supplémentaires à chaque ligne de données. Si vous désélectionnez cette case, seul le champ ID d'enregistrement et les champs de modélisation supplémentaires sont transmis et par conséquent, le flux s'exécute plus rapidement.

**Calcule les probabilités de classes affectées pour des enregistrements de scoring.** (Arbre de décision et Naive Bayes uniquement) Si cette option est sélectionnée, cela signifie que les champs de modélisation supplémentaires comprennent un champ de confiance (c'est-à-dire un champ de

probabilité) ainsi qu'un champ de prévision. Si vous désélectionnez cette case, seul le champ de prévision est généré.

- **Améliorez la précision des probabilités pour les ensembles de données petits ou très déséquilibrés.** Lors du calcul des probabilités, cette option appelle la technique de *m*-estimation pour éviter les probabilités zéro pendant l'estimation. Ce type d'estimation des probabilités peut être plus lent mais peut donner de meilleurs résultats pour les ensembles de données petits ou très déséquilibrés.

### ***Nuggets de modèles KNN Netezza***

Le nugget de modèle KNN permet de définir les options de scoring du modèle.

Quand vous exécutez un flux contenant un nœud de modélisation KNN, le nœud ajoute un nouveau champ, dont le nom est dérivé du nom du modèle.

Table 6-10  
*Champ de scoring des modèles pour KNN*

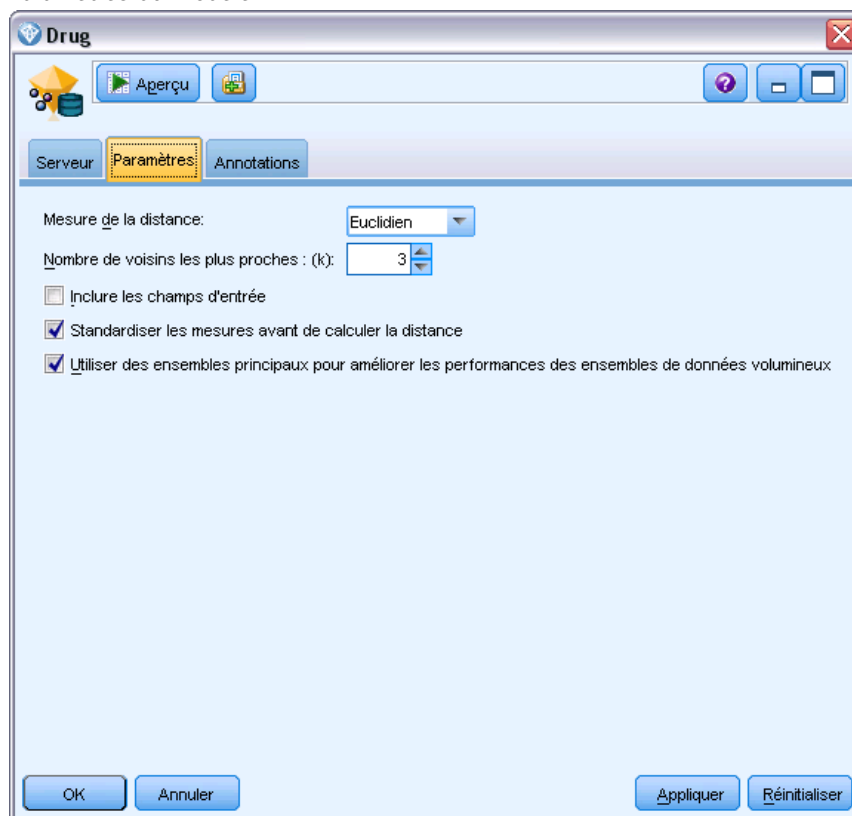
Nom du champ ajouté	Signification
\$KNN-nom_du_modèle	Valeur prédite pour l'enregistrement actuel.

Vous pouvez visualiser le champ supplémentaire en joignant un nœud Table au nugget de modèle et en exécutant le nœud Table. [Pour plus d'informations, reportez-vous à la section Détermination de scores des modèles IBM Netezza Analytics sur p. 220.](#)

### ***Nugget KNN Netezza - Onglet Paramètres***

Dans l'onglet Paramètres, vous pouvez définir les options pour le scoring du modèle.

Figure 6-36  
Paramètres du modèle KNN



**Mesure de distance.** La méthode à utiliser pour la mesure de distance entre les points de données ; plus les distances sont importantes plus les différences le sont aussi. Les options sont les suivantes :

- **Euclidienne.** (par défaut) La distance entre deux points est calculée en les reliant par une ligne droite.
- **Manhattan.** La distance entre deux points est calculée en tant que total des différences absolues entre leur coordonnées.
- **Canberra.** Semblable à la distance Manhattan, mais plus sensible aux points de données plus proches de l'origine.
- **Maximum :** la distance entre deux points est calculée comme la plus importante de leurs différences parmi toutes les dimensions de leurs coordonnées.

**Nombre de plus proches voisins (k).** Le nombre de plus proches voisins pour une observation particulière. Remarque : l'utilisation d'un nombre élevé de voisins ne garantit pas forcément un modèle plus précis.

Le choix de  $k$  contrôle l'équilibre entre la prévention du surajustement (ceci peut être important, en particulier pour les données parasites) et la résolution (donnant différentes prévisions pour des instances semblables). Vous devrez généralement régler la valeur de  $k$  pour chaque ensemble de données, avec des valeurs typiques allant de 1 à plusieurs douzaines.

**Inclure les champs d'entrée.** Si cette option est sélectionnée, elle transmet tous les champs d'entrée originaux en aval, ajoutant le ou les champs de modélisation supplémentaires à chaque ligne de données. Si vous désélectionnez cette case, seul le champ ID d'enregistrement et les champs de modélisation supplémentaires sont transmis et par conséquent, le flux s'exécute plus rapidement.

**Standardisez les mesures avant de calculer la distance.** Si elle est sélectionnée, cette option standardise les mesures des champs d'entrées continues avant de calculer les valeurs de distance.

**Utilisez des ensembles principaux pour améliorer les performances pour de grands ensembles de données.** Si elle est sélectionnée, cette option utilise l'échantillonnage d'ensembles principaux pour accélérer le calcul quand de grands ensembles de données sont impliqués.

### ***Nuggets des modèles de classification par division Netezza***

Le nugget de modèle de classification par division permet de définir les options de scoring du modèle.

Quand vous exécutez un flux contenant un nœud de modélisation de classification par division, le nœud ajoute deux nouveaux champs, dont les noms sont dérivés du nom du modèle.

Table 6-11  
*Champs de scoring du modèle pour la classification par division*

Nom du champ ajouté	Signification
\$DC-nom_du_modèle	Identificateur de la sous-classe à laquelle l'enregistrement actuel est attribué.
\$DCD-nom_du_modèle	Distance du centre de la sous-classe de l'enregistrement actuel.

Vous pouvez visualiser les champs supplémentaires en joignant un nœud Table au nugget de modèle et en exécutant le nœud Table. [Pour plus d'informations, reportez-vous à la section Détermination de scores des modèles IBM Netezza Analytics sur p. 220.](#)

### ***Nugget de classification par division Netezza - Onglet Paramètres***

Dans l'onglet Paramètres, vous pouvez définir les options pour le scoring du modèle.

**Inclure les champs d'entrée.** Si cette option est sélectionnée, elle transmet tous les champs d'entrée originaux en aval, ajoutant le ou les champs de modélisation supplémentaires à chaque ligne de données. Si vous désélectionnez cette case, seul le champ ID d'enregistrement et les champs de modélisation supplémentaires sont transmis et par conséquent, le flux s'exécute plus rapidement.

**Mesure de distance.** La méthode à utiliser pour la mesure de distance entre les points de données ; plus les distances sont importantes plus les différences le sont aussi. Les options sont les suivantes :

- **Euclidienne.** (par défaut) La distance entre deux points est calculée en les reliant par une ligne droite.
- **Manhattan.** La distance entre deux points est calculée en tant que total des différences absolues entre leur coordonnées.



- **Canberra.** Semblable à la distance Manhattan, mais plus sensible aux points de données plus proches de l'origine.
- **Maximum :** la distance entre deux points est calculée comme la plus importante de leurs différences parmi toutes les dimensions de leurs coordonnées.

**Niveau hiérarchique appliqué.** Le niveau hiérarchique devant être appliqué aux données.

## ***Nuggets de modèles ACP Netezza***

Le nugget de modèle ACP permet de définir les options de scoring du modèle.

Quand vous exécutez un flux contenant un nœud de modélisation ACP, le nœud ajoute un nouveau champ par défaut, dont le nom est dérivé du nom du modèle.

Table 6-12

*Champ de scoring du modèle pour l'ACP*

Nom du champ ajouté	Signification
\$IF-nom_du_modèle	Valeur prédite pour l'enregistrement actuel.

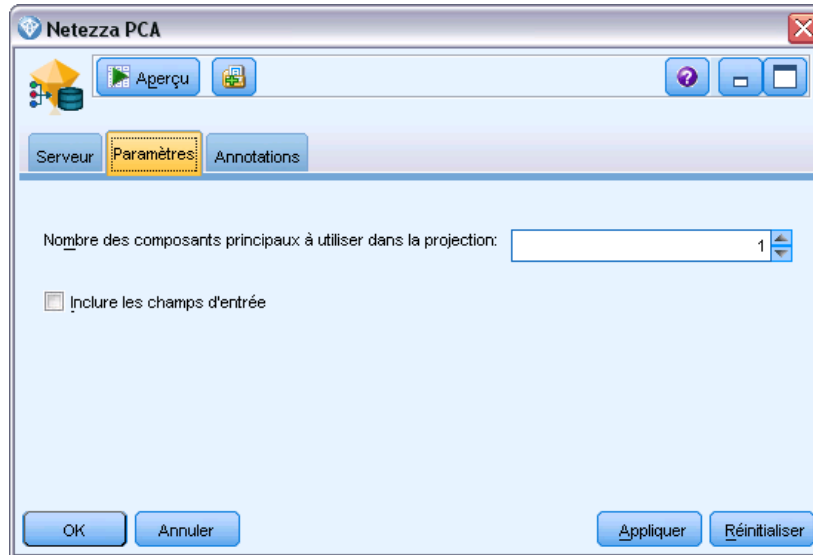
Si vous spécifiez une valeur supérieure à 1 dans le champ Nombre de composantes principales ... sur le nœud de modélisation ou le nugget de modélisation et exécutez le flux, le nœud ajoute un nouveau champ pour chaque composante. Dans ce cas, les noms de champs ont le suffixe *n*, où *n* est le numéro de la composante. Par exemple, si le nom de votre modèle est intitulé *acp* et si le modèle contient trois composantes, les noms des nouveaux champs sont *\$F-acp-1*, *\$F-acp-2* et *\$F-acp-3*.

Vous pouvez visualiser les champs supplémentaires en joignant un nœud Table au nugget de modèle et en exécutant le nœud Table. [Pour plus d'informations, reportez-vous à la section Détermination de scores des modèles IBM Netezza Analytics sur p. 220.](#)

### ***Nugget ACP Netezza - Onglet Paramètres***

Dans l'onglet Paramètres, vous pouvez définir les options pour le scoring du modèle.

Figure 6-37  
Paramètres du modèle ACP



**Nombre de composantes principales à utiliser dans la projection.** Le nombre de composantes principales auquel vous voulez réduire l'ensemble de données. Cette valeur ne doit pas dépasser le nombre d'attributs (champs d'entrée).

**Inclure les champs d'entrée.** Si cette option est sélectionnée, elle transmet tous les champs d'entrée originaux en aval, ajoutant le ou les champs de modélisation supplémentaires à chaque ligne de données. Si vous désélectionnez cette case, seul le champ ID d'enregistrement et les champs de modélisation supplémentaires sont transmis et par conséquent, le flux s'exécute plus rapidement.

## Nuggets de modèle Arbre de régression Netezza

Le nugget de modèle d'arbre de régression permet de définir les options de scoring du modèle.

Quand vous exécutez un flux contenant un nœud de modélisation Arbre de régression, le nœud ajoute un nouveau champ par défaut, dont le nom est dérivé du nom du modèle.

Table 6-13  
Champ de scoring des modèles pour l'arbre de régression

Nom du champ ajouté	Signification
\$I-nom_du_modèle	Valeur prédite pour l'enregistrement actuel.

Si vous avez sélectionné l'option Calculer la variance estimée sur le nœud de modélisation ou le nugget du modèle et exécuté le flux, un champ supplémentaire est ajouté.

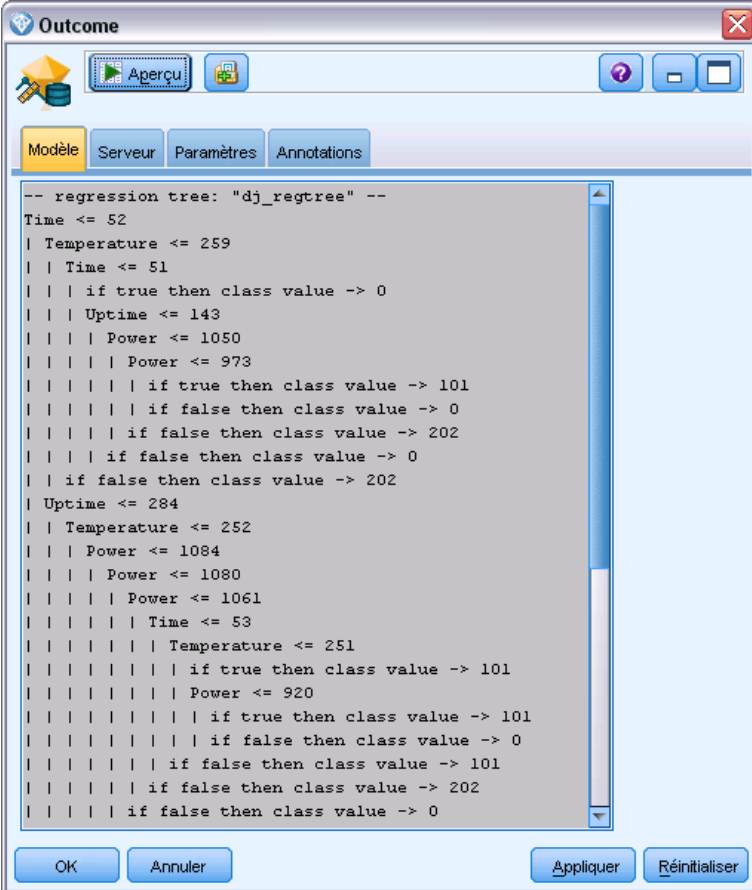
Table 6-14  
Champ de scoring des modèles pour l'arbre de régression - supplémentaire

Nom du champ ajouté	Signification
\$IV-nom_du_modèle	Variances estimées des classes attribuées.

Vous pouvez visualiser les champs supplémentaires en joignant un nœud Table au nugget de modèle et en exécutant le nœud Table. [Pour plus d'informations, reportez-vous à la section Détermination de scores des modèles IBM Netezza Analytics sur p. 220.](#)

### Nugget Arbre de régression Netezza - Onglet Modèle

Figure 6-38  
Sortie du modèle d'arbre de régression



```

-- regression tree: "dj_regtree" --
Time <= 52
| Temperature <= 259
| | Time <= 51
| | | if true then class value -> 0
| | | Uptime <= 143
| | | | Power <= 1050
| | | | | Power <= 973
| | | | | if true then class value -> 101
| | | | | if false then class value -> 0
| | | | | if false then class value -> 202
| | | | | if false then class value -> 0
| | | if false then class value -> 202
| | Uptime <= 284
| | | Temperature <= 252
| | | | Power <= 1084
| | | | | Power <= 1080
| | | | | | Power <= 1061
| | | | | | | Time <= 53
| | | | | | | | Temperature <= 251
| | | | | | | | if true then class value -> 101
| | | | | | | | | Power <= 920
| | | | | | | | | if true then class value -> 101
| | | | | | | | | if false then class value -> 0
| | | | | | | | | if false then class value -> 101
| | | | | | | | | if false then class value -> 202
| | | | | | | | | if false then class value -> 0

```

La sortie du modèle prend la forme d'une représentation textuelle de l'arbre. Chaque ligne de texte correspond à un nœud ou à une feuille, et la marge indique le niveau de l'arbre. Pour un nœud, la condition de division est affichée; pour une feuille, l'étiquette de classe attribuée apparaît.

### Nugget Arbre de régression Netezza - Onglet Paramètres

Dans l'onglet Paramètres, vous pouvez définir les options pour le scoring du modèle.

**Inclure les champs d'entrée.** Si cette option est sélectionnée, elle transmet tous les champs d'entrée originaux en aval, ajoutant le ou les champs de modélisation supplémentaires à chaque ligne de données. Si vous désélectionnez cette case, seul le champ ID d'enregistrement et les champs de modélisation supplémentaires sont transmis et par conséquent, le flux s'exécute plus rapidement.

**Calculer la variance estimée.** Indique si les variances des classes attribuées doivent être incluses dans la sortie.

### ***Nuggets du modèle de régression linéaire Netezza***

Le nugget de modèle de régression linéaire permet de définir les options de scoring du modèle.

Quand vous exécutez un flux contenant un nœud de modélisation de régression linéaire, le nœud ajoute un nouveau champ, dont le nom est dérivé du nom du modèle.

Table 6-15

*Champ de scoring des modèles pour la régression linéaire*

Nom du champ ajouté	Signification
\$LR-nom_du_modèle	Valeur prédite pour l'enregistrement actuel.

### ***Nugget de régression linéaire Netezza - Onglet Paramètres***

Dans l'onglet Paramètres, vous pouvez définir les options pour le scoring du modèle.

**Inclure les champs d'entrée.** Si cette option est sélectionnée, elle transmet tous les champs d'entrée originaux en aval, ajoutant le ou les champs de modélisation supplémentaires à chaque ligne de données. Si vous désélectionnez cette case, seul le champ ID d'enregistrement et les champs de modélisation supplémentaires sont transmis et par conséquent, le flux s'exécute plus rapidement.

### ***Nugget du modèle Séries temporelles Netezza***

Le nugget de modèle donne accès à la sortie de l'opération de modélisation des séries temporelles. La sortie est composée des champs suivants.

Table 6-16

*Champs de sortie de modèle de séries temporelles*

Champ	Description
TSID	L'identifiant des séries temporelles ; le contenu du champ spécifié pour les ID de séries temporelles dans l'onglet Champs du nœud de modélisation. <a href="#">Pour plus d'informations, reportez-vous à la section Options des champs de séries temporelles Netezza sur p. 206.</a>
TIME	La période dans les séries temporelles actuelles.
HISTORY	Les valeurs de données historiques (celles utilisées pour effectuer la prévision). Ce champ est inclus uniquement si l'option Inclure les valeurs historiques dans le résultat est sélectionnée dans l'onglet Paramètres du nugget de modèle.
\$STS-INTERPOLATED	Les valeurs interpolées lorsqu'elles sont utilisées. Ce champ est inclus uniquement si l'option Inclure les valeurs interpolées dans le résultat est sélectionnée dans l'onglet Paramètres du nugget de modèle. L'interpolation est une option de l'onglet Options de création du nœud de modélisation.
\$STS-FORECAST	Les valeurs de prévision des séries temporelles.

Pour afficher la sortie du modèle, joignez un nœud Table (de l'onglet Sortie de la palette de nœuds) au nugget de modèle et exécutez le nœud Table. Une sortie typique ressemble à la sortie suivante.

Figure 6-39  
Sortie typique du modèle Séries temporelles

	TSID	TIME	HISTORY	\$TS-INTERPOLATED	\$TS-FORECAST
22	m	1959-11-02	\$null\$	9.810	\$null\$
23	m	1960-07-17	15.000	\$null\$	\$null\$
24	m	1961-05-20	\$null\$	19.591	\$null\$
25	m	1962-07-18	15.000	\$null\$	\$null\$
26	m	1962-08-29	12.000	\$null\$	\$null\$
27	m	1962-12-07	\$null\$	3.401	\$null\$
28	m	1964-06-25	\$null\$	5.399	\$null\$
29	m	1964-11-17	12.000	\$null\$	\$null\$
30	m	1966-01-11	8.000	\$null\$	\$null\$
31	m	1967-07-31	\$null\$	\$null\$	0.590
32	m	1969-02-16	\$null\$	\$null\$	0.719
33	m	1970-09-04	\$null\$	\$null\$	0.667
34	m	1972-03-23	\$null\$	\$null\$	0.619
35	m	1973-10-10	\$null\$	\$null\$	0.574
36	m	1975-04-28	\$null\$	\$null\$	0.532
37	m	1976-11-14	\$null\$	\$null\$	0.494
38	m	1978-06-03	\$null\$	\$null\$	0.458
39	m	1979-12-20	\$null\$	\$null\$	0.425
40	m	1981-07-08	\$null\$	\$null\$	0.394
41	m	1983-01-25	\$null\$	\$null\$	0.366

### ***Nugget de séries temporelles Netezza - Onglet Paramètres***

Dans l'onglet Paramètres, vous pouvez spécifier les options de personnalisation de la sortie du modèle.

**Nom du modèle.** Le nom du modèle comme spécifié dans l'onglet Options du modèle du nœud de modélisation.

Les autres options sont les mêmes que celles de l'onglet Options de modélisation du nœud de modélisation.

### ***Nugget de modèle linéaire généralisé Netezza***

Le nugget de modèle donne accès à la sortie de l'opération de modélisation.

Lorsque vous exécutez un flux contenant un nœud de modélisation Linéaire généralisé, le nœud ajoute un nouveau champ dont le nom est dérivé du nom du modèle.

Table 6-17

Champ de scoring du modèle pour Linéaire généralisé

Nom du champ ajouté	Signification
\$GLM-model_name	Valeur prédite pour l'enregistrement actuel.

L'onglet Modèle affiche différentes statistiques associées au modèle.

Figure 6-40  
Sortie du modèle linéaire généralisé

Paramater	Beta	Std Error	Test
INTERCEPT	-3.514524	0	0
[ACE=15]	-0.014598	0	0
[ACE=16]	-0.089267	0	0
[ACE=17]	-0.120394	0	0
[ACE=18]	-0.273435	0	0
[ACE=19]	-0.421141	0	0
[ACE=20]	-0.10914	0	0
[ACE=21]	-0.463587	0	0
[ACE=22]	-0.173253	0	0
[ACE=23]	-0.173005	0	0
[ACE=24]	-0.097519	0	0
[ACE=25]	-0.067772	0	0
[ACE=26]	-0.226192	0	0
[ACE=28]	-0.195697	0	0
[ACE=29]	-0.323554	0	0
[ACE=30]	-0.342686	0	0
[ACE=31]	-0.202401	0	0
[ACE=32]	-0.063812	0	0
[ACE=33]	-0.638801	0	0
[ACE=34]	-0.086032	0	0
[ACE=35]	-0.018557	0	0
[ACE=36]	-0.123014	0	0
[ACE=37]	-0.080038	0	0
[ACE=38]	-0.216761	0	0

La sortie est composée des champs suivants.

Table 6-18  
Champs de sortie du modèle linéaire généralisé

Champ de sortie	Description
Paramètre	Les paramètres (c'est-à-dire les variables prédictives) utilisés par le modèle. Il s'agit des colonnes numériques et nominales ainsi que de la constante (celle du modèle de régression).
Bêta	Le coefficient de corrélation (c'est-à-dire le composant linéaire du modèle).
Erreur-type	L'écart-type pour la Bêta.
Test	Les statistiques de test utilisées pour évaluer la validité du paramètre.
valeur p	La probabilité d'une erreur lorsque que le paramètre est considéré comme significatif.
<b>Récapitulatif des résidus</b>	
Type de résidus	Le type de résidus de la prévision pour laquelle les valeurs récapitulatives apparaissent.
RSS	La valeur des résidus.
ddl	Les degrés de liberté des résidus.
valeur p	La probabilité d'une erreur. Une valeur élevée indique un modèle mal ajusté, une valeur faible indique un modèle bien ajusté.

***Nugget de modèle linéaire généralisé Netezza - Onglet Paramètres***

Dans l'onglet Paramètres, vous pouvez personnaliser la sortie du modèle.

Cette option est la même que celle des Options de scoring dans le nœud de modélisation. [Pour plus d'informations, reportez-vous à la section Options de modèle linéaire généralisé Netezza - Options de scoring sur p. 220.](#)

## Remarques

Ces informations ont été développées pour les produits et services offerts dans le monde.

Il est possible qu'IBM n'offre pas dans les autres pays les produits, services et fonctionnalités décrits dans ce document. Contactez votre représentant local IBM pour obtenir des informations sur les produits et services actuellement disponibles dans votre région. Toute référence à un produit, programme ou service IBM n'implique pas que les seuls les produits, programmes ou services IBM peuvent être utilisés. Tout produit, programme ou service de fonctionnalité équivalente qui ne viole pas la propriété intellectuelle IBM peut être utilisé à la place. Cependant l'utilisateur doit évaluer et vérifier l'utilisation d'un produit, programme ou service non IBM.

IBM peut posséder des brevets ou des applications de brevet en attente qui couvrent les sujets décrits dans ce document. L'octroi de ce document n'équivaut aucunement à celui d'une licence pour ces brevets. Vous pouvez envoyer par écrit des questions concernant la licence à :

*IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785, États-Unis*

Pour obtenir des informations de licence concernant la configuration de caractères codés sur deux octets (DBCS), veuillez contacter dans votre pays le département chargé de la propriété intellectuelle chez IBM ou envoyez vos commentaires par écrit à :

*Intellectual Property Licensing, Legal and Intellectual Property Law, IBM Japan Ltd., 1623-14, Shimotsuruma, Yamato-shi, Kanagawa 242-8502 Japon.*

**Le paragraphe suivant ne s'applique pas au Royaume-Uni ni à aucun pays dans lequel ces dispositions sont contraires au droit local :** INTERNATIONAL BUSINESS MACHINES FOURNIT CETTE PUBLICATION « EN L'ÉTAT » SANS GARANTIE D'AUCUNE SORTE, IMPLICITE OU EXPLICITE, Y COMPRIS, MAIS SANS ETRE LIMITE AUX GARANTIES IMPLICITES DE NON VIOLATION, DE QUALITE MARCHANDE OU D'ADAPTATION POUR UN USAGE PARTICULIER. Certains états n'autorisent pas l'exclusion de garanties explicites ou implicites lors de certaines transactions, par conséquent, il est possible que cet énoncé ne vous concerne pas.

Ces informations peuvent contenir des erreurs techniques ou des erreurs typographiques. Ces informations sont modifiées de temps en temps ; ces modifications seront intégrées aux nouvelles versions de la publication. IBM peut apporter des améliorations et/ou modifications des produits et/ou des programmes décrits dans cette publications à tout moment sans avertissement préalable.

Toute référence dans ces informations à des sites Web autres qu'IBM est fournie dans un but pratique uniquement et ne sert en aucun cas de recommandation pour ces sites Web. Le matériel contenu sur ces sites Web ne fait pas partie du matériel de ce produit IBM et l'utilisation de ces sites Web se fait à vos propres risques.

IBM peut utiliser ou distribuer les informations que vous lui fournissez, de la façon dont il le souhaite, sans encourir aucune obligation envers vous.



Les personnes disposant d'une licence pour ce programme et qui souhaitent obtenir des informations sur celui-ci pour activer : (i) l'échange d'informations entre des programmes créés de manière indépendante et d'autres programmes (notamment celui-ci) et (ii) l'utilisation mutuelle des informations qui ont été échangées, doivent contacter :

*IBM Software Group, Attention: Licensing, 233 S. Wacker Dr., Chicago, IL 60606, États-Unis.*

Ces informations peuvent être disponibles, soumises à des conditions générales, et dans certains cas payantes.

Le programme sous licence décrit dans ce document et toute la documentation sous licence disponible pour ce programme sont fournis par IBM en conformité avec les conditions de l'accord du client IBM, avec l'accord de licence du programme international IBM et avec tout accord équivalent entre nous.

Toutes les données sur les performances contenues dans le présent document ont été obtenues dans un environnement contrôlé. Par conséquent, les résultats obtenus dans d'autres environnements d'exploitation peuvent varier de manière significative. Certaines mesures peuvent avoir été effectuées sur des systèmes en cours de développement et il est impossible de garantir que ces mesures seront les mêmes sur les systèmes commercialisés. De plus, certaines mesures peuvent avoir été estimées par extrapolation. Les résultats réels peuvent être différents. Les utilisateurs de ce document doivent vérifier les données applicables à leur environnement spécifique.

Les informations concernant les produits autres qu'IBM ont été obtenues auprès des fabricants de ces produits, leurs annonces publiques ou d'autres sources publiques disponibles. IBM n'a pas testé ces produits et ne peut confirmer l'exactitude de leurs performances, leur compatibilité ou toute autre fonctionnalité associée à des produits autres qu'IBM. Les questions sur les capacités de produits autres qu'IBM doivent être adressées aux fabricants de ces produits.

Toutes les déclarations concernant la direction ou les intentions futures d'IBM peuvent être modifiées ou retirées sans avertissement préalable et représentent uniquement des buts et des objectifs.

Ces informations contiennent des exemples de données et de rapports utilisés au cours d'opérations quotidiennes standard. Pour les illustrer le mieux possible, ces exemples contiennent des noms d'individus, d'entreprises, de marques et de produits. Tous ces noms sont fictifs et toute ressemblance avec des noms et des adresses utilisés par une entreprise réelle ne serait que pure coïncidence.

Si vous consultez la version papier de ces informations, il est possible que certaines photographies et illustrations en couleurs n'apparaissent pas.

### **Marques commerciales**

IBM, le logo IBM, ibm.com et SPSS sont des marques commerciales d'IBM Corporation, déposées dans de nombreuses juridictions du monde entier. Une liste à jour des marques IBM est disponible sur Internet à l'adresse <http://www.ibm.com/legal/copytrade.shtml>.

Intel, le logo Intel, Intel Inside, le logo Intel Inside, Intel Centrino, le logo Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium, et Pentium sont des marques commerciales ou des marques déposées de Intel Corporation ou de ses filiales aux États-Unis et dans d'autres pays.

Linux est une marque déposée de Linus Torvalds aux États-Unis et/ou dans d'autres pays.

Microsoft, Windows, Windows NT et le logo Windows sont des marques commerciales de Microsoft Corporation aux Etats-Unis et/ou dans d'autres pays.

UNIX est une marque déposée de The Open Group aux Etats-Unis et dans d'autres pays.

Java et toutes les marques et logos Java sont des marques commerciales de Sun Microsystems, Inc. aux Etats-Unis et/ou dans d'autres pays.

Les autres noms de produits et de services peuvent être des marques d'IBM ou d'autres sociétés.



# Index

- A priori
  - Microsoft, 30
  - Oracle Data Mining, 84, 87
- Adaptive Bayes Network
  - Oracle Data Mining, 64–65, 67
- analyse spectrale, IBM Netezza Analytics, 204
- Analysis Services
  - Arbres décision, 47
  - Exemples, 47
  - gestion des modèles, 21
  - intégration avec IBM SPSS Modeler, 8
  - intégration dans IBM SPSS Modeler, 15
- Arbre de décision
  - IBM Netezza Analytics, 175, 177–178, 180–181, 222–223
  - Oracle Data Mining, 75–77
- arbres de régression
  - IBM Netezza Analytics, 199, 201, 232–233
- arbres décision
  - Microsoft Analysis Services, 14, 17, 38
  - options de modèle, 23
  - options du serveur, 23
  - options expert, 25
  - scoring - options du serveur, 39
  - scoring - options récapitulatives, 41
- base de données
  - modélisation dans la base de données pour ISW, 108
  - modélisation de base de données, 11, 14, 17, 20, 38
- Centrer-réduire
  - normalisation de données, 99
- champ unique
  - Algorithme Oracle k-Means, 80
  - Apriori Oracle, 76, 87
  - MDL Oracle, 89
  - NMF Oracle, 82
  - O-Cluster Oracle, 78
  - Oracle Adaptive Bayes Network, 66
  - Oracle Data Mining, 59
  - Oracle Naive Bayes, 63
  - Oracle Support Vector Machine, 68
- champs de partition
  - Sélection, 86
- Classification
  - IBM Netezza Analytics, 230
  - InfoSphere Warehouse Data Mining, 142
  - options de modèle, 23
  - options du serveur, 23
  - options expert, 26
  - scoring - options du serveur, 39
  - scoring - options récapitulatives, 41
- classification de séquences
  - options de modèle, 23
  - classification de séquences (Microsoft), 36
    - options de champs, 36
    - options expert, 38
  - classification par division
    - IBM Netezza Analytics, 193, 195
  - Classification par division
    - IBM Netezza Analytics, 230
  - clé
    - clés de modèle, 13
  - coûts
    - Oracle, 61
  - coûts de mauvaise réaffectation
    - arbres de décision, 61, 122
    - Oracle, 61
  - critère de division
    - Algorithme Oracle k-Means, 80
  - cross-validation (validation croisée)
    - Oracle Naive Bayes, 62
- DB2
  - gestion des modèles, 117
- décomposition des tendances saisonnières, IBM Netezza Analytics, 204
- déploiement, 53, 106, 163
- Description de longueur minimale (MDL)
  - Oracle Data Mining, 88–89
- détermination de scores du modèle
  - InfoSphere Warehouse Data Mining, 116
- discrétisation des données
  - modèles Oracle, 99
- documentation, 4
- données tabulaires
  - Noeud Association ISW, 127
- données transactionnelles
  - Noeud Association ISW, 127
- DSN
  - configuration, 17
- Ecart  $z$ 
  - normalisation de données, 69
- Ecart-type
  - Oracle Support Vector Machine, 70
- éditeur de catégorie
  - Noeud Association ISW, 132
- epsilon
  - Oracle Support Vector Machine, 70
- étiquette de classe, dans les modèles d'arbre Netezza, 175
- évaluation, 50, 103, 161
- exemples
  - Aperçu, 6
- Exemples
  - exploration de base de données, 46–50, 53, 101, 158–161, 163
  - Guide des applications, 4

- exemples d'application, 4
- exploration, 48, 101, 159
- exploration de base de données
  - configuration, 17
  - création de modèles, 10
  - exemple, 46, 158
  - options d'optimisation, 11
  - préparation des données, 11
  - utilisation IBM SPSS Modeler, 9
- exportation
  - modèles Analysis Services, 46
  - modèles DB2, 119
- facteur de complexité
  - Oracle Support Vector Machine, 70
- feuille, dans les modèles d'arbre Netezza, 175
- fichier *tnsnames.ora*, 57
- flux
  - Exemples d'InfoSphere Warehouse Data Mining, 158
- fonction de distance
  - Algorithme Oracle k-Means, 80
- génération de noeuds, 46
- génération SQL, 8, 11
- IBM
  - gestion des modèles, 117
  - Modélisation d'arbres décision, 107
  - Modélisation d'associations, 107
  - Modélisation de classification non supervisée Kohonen, 107
  - Modélisation de classifications démographiques, 107
  - Modélisation de régression linéaire, 107
  - Modélisation de régression logistique, 107
  - Modélisation de régression polynomiale, 107
  - Modélisation de régressions, 107
  - Modélisation de séquences, 107
  - Modélisation de séries temporelles, 107
  - Modélisation Naive Bayes, 107
- IBM InfoSphere Warehouse (ISW)
  - intégration avec IBM SPSS Modeler, 8
- IBM Netezza Analytics, 165
  - ACP, 196
  - Agrégation suivant le saut minimum (KNN), 189
  - Arbre de régression, 199
  - Arbres décision, 175
  - Bayes Net, 186
  - Classification par division, 193
  - configuration avec IBM SPSS Modeler, 165–166, 170, 172
  - gestion des modèles, 220–221
  - K-means, 183
  - Linéaire généralisé, 215
  - Naive Bayes, 188
  - Nugget de modèle ACP, 231
  - Nugget de modèle Bayes Net, 225–226
  - Nugget de modèle de nuées dynamiques, 224–225
  - Nugget de modèle KNN, 228
  - Nugget de modèle linéaire généralisé, 235, 237
  - Nugget des modèles Naive Bayes, 227
  - Nugget du modèle d'arbre de régression, 232–233
  - Nugget du modèle de classification par division, 230
  - Nugget du modèle de régression linéaire, 234
  - nugget du modèle Séries temporelles, 234–235
  - Nuggets de modèle Arbre de décision, 222–223
  - Options de champ Classification par division, 193
  - options de champs, 171
  - Options de champs ACP, 196
  - Options de champs Bayes Net, 186
  - Options de création ACP, 198
  - Options de création Bayes Net, 187
  - Options de création Classification par division, 195
  - Options de création d'arbre décision, 178, 180–181
  - Options de création d'arbre de régression, 199, 201
  - Options de création de séries temporelles, 208, 212
  - Options de création K-Means, 185
  - Options de création Régression linéaire, 202
  - options de modèle, 174
  - Options de modèle de séries temporelles, 213
  - Options de modèle KNN, 189, 191
  - Options de modèle linéaire généralisé, 215, 217
  - Options des champs de séries temporelles, 206
  - Options du champ Arbre décision, 177
  - Options du champ K-Means, 183
  - Régression linéaire, 202
  - Séries temporelles, 204
- IBM SPSS Modeler, 1
  - documentation, 4
  - exploration de base de données, 9
- IBM SPSS Modeler Solution Publisher
  - modèles Oracle Data Mining, 60
- Importance de l'attribut (IA)
  - Oracle Data Mining, 90–92
- InfoSphere Warehouse (IBM), voir ISW, 108
- InfoSphere Warehouse Data Mining
  - arbres décision, 123
  - exemples de flux, 158
  - Modélisation d'associations, 125
  - Noeud Régression, 137
  - Noeud Séquence, 134
  - nuggets de modèle, 155
  - taxonomie, 131
- interpolation des valeurs, séries temporelles IBM Netezza Analytics, 205
- ISW
  - connexion ODBC, 108
  - intégration avec IBM SPSS Modeler, 108
  - Onglet Serveur, 120
- k-Means
  - IBM Netezza Analytics, 183, 185
  - Oracle Data Mining, 79–81

- K-means
  - IBM Netezza Analytics, 224–225
- lissage exponentiel
  - IBM Netezza Analytics, 204
- marques commerciales, 239
- MDL, 64
- mentions légales, 238
- mesure d'impureté Entropy, 179
- mesure d'impureté Gini, 179
- mesures d'impureté
  - Arbre décision Netezza, 179
- méthode de normalisation
  - Algorithme Oracle k-Means, 80
  - NMF Oracle, 82
  - Oracle Support Vector Machine, 69
- Microsoft
  - Analysis Services, 14, 17, 38
  - Classification de séquences, 14
  - gestion des modèles, 21
  - Modélisation d'arbres décision, 14, 17, 38
  - Modélisation de classification non supervisée, 14, 17, 38
  - Modélisation de règles d'association, 14, 17, 38
  - Modélisation de régression linéaire, 17, 38
  - Modélisation de régression logistique, 17, 38
  - Modélisation de réseau de neurones, 17, 38
  - Modélisation Naive Bayes, 14, 17, 38
  - Régression linéaire, 14
  - Régression logistique, 14
  - Réseau de neurones, 14
- Microsoft Analysis Services, 42, 45–46
  - intégration dans IBM SPSS Modeler, 15
- Microsoft SQL Server
  - intégration dans IBM SPSS Modeler, 15
- Min-Max
  - normalisation de données, 69, 99
- Minimum Description Length, 64
- Modèles
  - création de modèles dans une base de données, 10
  - enregistrement, 12
  - évaluation, 50, 103, 161
  - exploration dans Oracle, 65
  - Exportation, 12
  - gestion d'Analysis Services, 21
  - gestion de DB2, 117
  - liste DB2, 118
  - navigation dans DB2, 119
  - problèmes de cohérence, 13
  - scoring de modèles dans une base de données, 11
- modèles à fonction unique
  - Oracle Adaptive Bayes Network, 66
- modèles ACP
  - IBM Netezza Analytics, 196, 198, 231
- modèles ARIMA
  - IBM Netezza Analytics, 204, 211
- modèles d'agrégation suivant le saut minimum
  - IBM Netezza Analytics, 189, 191, 228
- modèles d'arbre décision
  - InfoSphere Warehouse Data Mining, 123
- modèles de règles d'association
  - Microsoft, 30
- modèles de réseau Bayésien
  - IBM Netezza Analytics, 186–187, 225–226
- modèles KNN
  - IBM Netezza Analytics, 228
- modèles linéaires généralisés
  - IBM Netezza Analytics, 215, 217, 219–220, 235, 237
- Modèles linéaires généralisés (MLG)
  - Oracle Data Mining, 72–74
- modèles multifonctions
  - Oracle Adaptive Bayes Network, 66
- modèles Naive Bayes
  - IBM Netezza Analytics, 227
  - Oracle Adaptive Bayes Network, 66
- modèles Naive Bayes élagués
  - Oracle Adaptive Bayes Network, 66
- Modélisation d'associations
  - InfoSphere Warehouse Data Mining, 125
- modélisation de base de données, 41
  - Analysis Services, 8
  - IBM InfoSphere Warehouse (ISW), 8
  - Oracle Data Miner, 8
- modélisation de bases de données
  - Analysis Services, 8
  - IBM InfoSphere Warehouse (ISW), 8
  - IBM Netezza Analytics, 165–166, 170, 172
  - Oracle, 55–56, 59–60
  - Oracle Data Miner, 8
- Naive Bayes
  - IBM Netezza Analytics, 188, 227
  - InfoSphere Warehouse Data Mining, 148
  - options de modèle, 23
  - options du serveur, 23
  - options expert, 27
  - Oracle Data Mining, 62–64
  - scoring - options du serveur, 39
  - scoring - options récapitulatives, 41
- NMF
  - Oracle Data Mining, 82–83
- nœud Audit données, 48, 101, 159
- Noeud Classification
  - InfoSphere Warehouse Data Mining, 142
- Noeud Publisher
  - modèles Oracle Data Mining, 60
- Noeud Régression
  - InfoSphere Warehouse Data Mining, 137
- noeud Régression logistique
  - InfoSphere Warehouse Data Mining, 149
- Noeud Séquence
  - InfoSphere Warehouse Data Mining, 134

- noeuds
  - génération, 46
- noeuds de modélisation
  - Arbre décision MS, 20
  - Classification de séquences Microsoft, 20
  - classification non supervisée MS, 20
  - modélisation dans la base de données pour ISW, 108
  - modélisation de base de données, 11, 14, 17, 20, 38
  - Naive Bayes MS, 20
  - règles d'association Microsoft, 20
  - Régression linéaire Microsoft, 20
  - Régression logistique Microsoft, 20
  - Réseau de neurones Microsoft, 20
  - Séries temporelles Microsoft, 20
- nom d'hôte
  - connexion Oracle, 57
- Nombre de classes
  - Algorithme Oracle k-Means, 80
  - O-Cluster Oracle, 78
- normalisation de données
  - modèles Oracle, 99
- noyau gaussien
  - Oracle Support Vector Machine, 67
- noyau linéaire
  - Oracle Support Vector Machine, 67
- nuggets de modèle
  - IBM Netezza Analytics, 222–228, 230–235, 237
  - InfoSphere Warehouse Data Mining, 155
- O-Cluster
  - Oracle Data Mining, 78–79
- ODBC
  - configuration, 17
  - configuration de IBM Netezza Analytics, 165–166, 170, 172
  - configuration de SQL Server, 18
  - configuration d'ISW, 108
  - configuration pour Oracle, 55–56, 59–60
- ODM. *Voir* Oracle Data Mining, 55
- Onglet Serveur
  - ISW, 120
- optimisation
  - génération SQL, 8
- Optimisation SQL. *Voir* Génération SQL, 8
- options de champs
  - IBM Netezza Analytics, 171, 177, 183, 186, 193, 196, 198, 206
  - noeuds de modélisation, 126
- options de création
  - IBM Netezza Analytics, 178, 180–181, 185, 187, 195, 199, 201–202, 208, 212
- options de modèle
  - IBM Netezza Analytics, 174, 189, 191, 213, 215, 217
- options de puissance
  - ISW Data Mining, 121
- Oracle Data Miner, 97
  - intégration avec IBM SPSS Modeler, 8
- Oracle Data Mining, 55
  - A priori, 84, 87
  - Adaptive Bayes Network, 64–65, 67
  - Arbre de décision, 75–77
  - configuration avec IBM SPSS Modeler, 55–56, 59–60
  - coûts de mauvaise réaffectation, 95
  - Description de longueur minimale (MDL), 88–89
  - Exemples, 99–103, 106
  - gestion des modèles, 93–95
  - Importance de l'attribut (IA), 90–92
  - k-Means, 79–81
  - Modèles linéaires généralisés (MLG), 72–74
  - Naive Bayes, 62–64
  - NMF, 82–83
  - O-Cluster, 78–79
  - préparation des données, 99
  - Support Vector Machine, 67–69
  - vérification de la cohérence, 93
- partition des données, 86
- partitions, 127
  - création de modèles, 33, 63, 66, 90, 129, 135, 139, 144, 149–150
  - Sélection, 127
- pénalité pour complexité, 25–31, 34
- pondération de classe, dans les modèles d'arbre Netezza, 176
- pondération d'instance, dans les modèles d'arbre Netezza, 176
- port
  - connexion Oracle, 57
- Probabilités a priori
  - Oracle Data Mining, 71
- règles d'association
  - options de modèle, 23
  - options du serveur, 23
  - options expert, 31
  - scoring - options du serveur, 39
  - scoring - options récapitulatives, 41
- régression linéaire
  - IBM Netezza Analytics, 199, 202, 234
  - options de modèle, 23
  - options du serveur, 23
  - options expert, 28
  - scoring - options du serveur, 39
  - scoring - options récapitulatives, 41
- Régression logistique
  - options de modèle, 23
  - options du serveur, 23
  - options expert, 30
  - scoring - options du serveur, 39
  - scoring - options récapitulatives, 41
- Répercussions SQL. *Voir* Génération SQL, 8
- réseau de neurones
  - options de modèle, 23

- options du serveur, 23
- options expert, 29
- scoring - options du serveur, 39
- scoring - options récapitulatives, 41
  
- scoring, 11, 220
- Séries temporelles
  - IBM Netezza Analytics, 206, 208, 212–213
  - InfoSphere Warehouse Data Mining, 150–154
- séries temporelles (IBM Netezza Analytics), 234–235
- Séries temporelles (IBM Netezza Analytics), 204
- séries temporelles (Microsoft), 32
  - options de modèle, 33
  - options de paramètres, 35
  - options expert, 34
- serveur
  - exécution d'Analysis Services, 23, 39, 41
- seuil de singleton
  - Oracle Naive Bayes, 64
- seuil par paire
  - Oracle Naive Bayes, 64
- SID
  - connexion Oracle, 57
- Solution Publisher
  - modèles Oracle Data Mining, 60
- SPSS Modeler Server, 2
- SQL Server, 23, 39, 41
  - configuration, 17
  - connexion ODBC, 18
  - intégration dans IBM SPSS Modeler, 15
- Support Vector Machine
  - Oracle Data Mining, 67–69
- SVM. *Voir* Support Vector Machine, 67
  
- taxonomie
  - InfoSphere Warehouse Data Mining, 131
- tolérance de convergence
  - Oracle Support Vector Machine, 69
  
- valeurs numériques d'impureté
  - Apriori Oracle, 76